

Astronomy of the Brain: Neural Reorganization Across Persistent Fentanyl Use

Ari Mendel Peden-Asarch

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

University of Washington
2026

Reading Committee:
John F. Neumaier, Chair
Kevin R. Coffey
Sheri J.Y. Mizumori

Program Authorized to Offer Degree:
Neuroscience

© Copyright 2026

Ari Mendel Peden-Asarch

University of Washington

Abstract

Astronomy of the Brain: Neural Reorganization Across Persistent Fentanyl Use

Ari Mendel Peden-Asarch

Chair of the Supervisory Committee:

John F. Neumaier, M.D., Ph.D.

Department of Psychiatry & Behavioral Sciences

Persistent opioid use despite adverse consequences involves a series of events that unfold while the drug is already on board, yet preclinical models usually investigate choices made drug-free. To address this, we developed a delayed shock paradigm for oral fentanyl self-administration in rats. After a stable baseline of fentanyl intake was established, we escalated shock intensity across days (0.1, 0.2, 0.4, and 0.8 mA), spanning the mild intensities conventional in self-administration work and the higher intensities conventional in stress work, and then imposed a week of forced abstinence followed by a Relapse test at 0.8 mA. During the shock sessions the animals could reach their preferred fentanyl concentration during the Safe period (15 trials), which was followed by the Danger period (remaining 165 trials) in which any rewarded lever press delivered the reward and then a shock 4 s later. We found that the Danger-trial responding was indistinguishable from baseline through 0.4 mA and collapsed at 0.8 mA, so the same animals would have been called shock resistant by the self-administration convention and shock sensitive by the stress convention, and both readings would have been artifacts of an inherited intensity rather than facts about the animal. A within-session changepoint model showed that shock moved the timing of the strategy switch rather than whether it occurred, and ultrasonic vocalizations reorganized across the same ladder while dissociating from responding, so the switch was also captured by a measure that did not depend on lever pressing. This paradigm demanded multi-hour recordings and the ability to register the same neurons across weeks, and neither capability existed. We therefore built: 1) The Miniscope Processing Suite (MPS), which carried 7.26 TB and 77 h of raw video through preprocessing, source extraction, and deconvolution in 55.6 h on a single workstation and recovered every neuron of a matched synthetic ground truth, and 2) Stars2Cells (S2C), which registers neurons from the geometry of their four-neuron neighborhoods, reaching an F1 of 98.4% across a 1,262-run synthetic benchmark, reproducing the Allen Brain Observatory registration across 431 two-photon containers at a median F1 of 0.97, and retaining $41.6 \pm 3.4\%$ of first-session neurons across a four-session chain against $13.0 \pm 3.9\%$ for footprint matching. We validated this tool stack in the dorsomedial striatum (DMS) during a fentanyl behavioral-economics task. Same-cell tracking revealed that the press-responsive fraction held near one half across the demand curve while roughly half of the

tracked neurons switched class at every transition. Applying the same tools to the lateral habenula (LHb) in our aforementioned paradigm, we found that the shock was the most decodable event in the session (ROC-AUC 0.88 at 0.8 mA) and that a shock ensemble of comparable strength was recovered at Relapse a week later, yet the population reported a press only after it had occurred and forecast neither an upcoming press nor the switch to withholding. Thus, we establish that the LHb registers what a choice cost rather than computing what to do next under adverse consequences during active fentanyl self-administration. Furthermore, since a policy for stopping was acquired while the drug was on board and remained stable across weeks, we suggest that treatment needs to address the brain state under active drug use and not only during abstinence.

For Nicki, who believed in me, and in us.

Acknowledgments

To John, for teaching me how to be a good scientist, a good mentor, and a good person, and for the opportunity itself. I am honored to have been your student. To Kevin, for his patience in teaching me more things than I can count, for his understanding, and for the confidence to build methods rather than wait for them. To Sheri, for helping me feel that I belonged. And to Chet, for his support and for the opening that started my work in machine learning and code.

To Michele, our lab manager, without whom this would not exist, and for whom I have the deepest respect. To Nathan, for teaching me how to navigate an institution. To Will, who showed me the ropes. To Jenna and Madelyn, the best post-baccs and now lab techs a lab could hope for. To Jackie, whose energy is infectious. Before any of it, to Paul and Lauren, for starting me down this path. To Ryan and Lydia, for teaching me how to work independently. And to Jacqueline, for teaching me how to have fun in science. To Jennifer, Natalia, Céline, and Kyle, for their unwavering and tremendous support. And to the University of Washington, where I learned philosophy, neuroscience, and then how to do it myself (Go Dawgs!).

To my mother Jodi, for the effort she put in. To my father Chad, for what he gave me that I chose to keep. To my brother Eli, who taught me what I wanted out of this. To my sister Emi, who made me feel like I mattered. To Wendy and Dave, who taught me how to build a home. To Jenny and Troy, who taught me how to let people in. To Neal and Katie, for their unwavering support. To my grandparents Rich and Elaine, who taught me how to be an artist. To Lee, my grandfather, who taught me how to swing easy. And to Audi, my dog, who was a very good boy.

To all of my friends, for everything and for the good times.

And to Nicki, for always being my home.

Funding: NIDA T32DA007278, R01DA052618, R00DA052571; Puget Sound VA Research and Development Pilot Grant.

Contents

Acknowledgments	v
1 General Introduction	1
1.1 The opioid crisis and decisions made while the drug is on board	1
1.2 The lateral habenula as an outcome signal	2
1.3 The methodological bottleneck of scale and identity across sessions	3
1.4 Tools, terms, and the theories they carry	4
1.5 Dissertation aims	6
2 A No-code and Improved Pipeline for Long-Duration Miniscope Calcium Imaging	7
2.1 Introduction	7
2.2 Results	9
2.2.1 Pipeline Overview & Installation (Figure 1: Step 1)	9
2.2.2 Video Preprocessing (Figure 2: Steps 2a-2f)	13
2.2.3 ROI Cropping & Dimensionality Reduction (Figure 3: Steps 3a-3c)	14
2.2.4 Cell detection, Initialization, and Pre-CNMF Validation (Figure 4: Steps 4a-5b)	16
2.2.5 Temporal update (Figure 5: Steps 6a-6e)	18
2.2.6 Spatial update (Figure 6: Steps 7a-7f)	20
2.2.7 Final Temporal Update (Figure 7: Steps 8a-8c)	22
2.2.8 Validation Against Synthetic Ground Truth (Figure 8)	24
2.2.9 Summary	25
2.3 Discussion	25
2.3.1 Accessibility	25
2.3.2 Scalability for Longer Recordings	25
2.3.3 Extensibility to New Modalities	26
2.3.4 Limitations	26
2.3.5 Community Impact	26
2.4 Methods and Materials	27
2.4.1 Dataset Description and Scope of Validation	27
2.4.2 Synthetic Ground-Truth Validation Dataset	27
2.4.3 Ground-Truth Recovery Analysis	28

2.4.4	Step 0: Download	29
2.4.5	Step 1: Graphical User Interface, Environment Provisioning, and Automation	31
2.4.6	Step 2: Data Preprocessing	32
2.4.7	Step 3: Field-of-View Cropping and Dimensionality Reduction	36
2.4.8	Step 4: Initial Cell Detection and Component Initialization	37
2.4.9	Step 5: Pre-CNMF Noise Estimation and Validation	40
2.4.10	Step 6: Temporal Activity Update (CNMF Temporal Refinement)	41
2.4.11	Step 7: Spatial Footprint Update (CNMF Spatial Refinement)	43
2.4.12	Step 8: Final Temporal Refinement and Output	45
2.4.13	Interactive Data Explorer and Manual Curation	46
2.4.14	Parallel Computing Implementation and Scalability Considerations	47
2.4.15	Statistical Analyses	48
2.4.16	Citations	50
2.5	Acknowledgments	50
3	Astrometric Tracking of Cells Across Imaging Sessions	51
3.1	Introduction	51
3.2	Results	55
3.2.1	Quad descriptor pairing from neuron centroid constellations	55
3.2.2	Activity correlation alone cannot identify neurons across sessions	58
3.2.3	Synthetic validation across recording modalities	60
3.2.4	Longitudinal DMS tracking across a fentanyl behavioral economics paradigm	63
3.2.5	Behavioral demand and longitudinal DMS population dynamics	65
3.3	Discussion	68
3.4	Methods	71
3.4.1	Stars2Cells algorithm	71
3.4.2	Temporal correlation control experiment	74
3.4.3	Synthetic benchmark generation	75
3.4.4	Spatial footprint synthesis for the ROI-based matching comparison	75
3.4.5	ROI-based matching reimplementatation	76
3.4.6	Evaluation metrics	79
3.4.7	Bibliometric adoption analysis	79
3.4.8	Cross-modality validation on the Allen Brain Observatory	80
3.4.9	Animals and housing	80
3.4.10	Oral fentanyl self-administration and behavioral economics	81
3.4.11	Pose estimation and kinematic analysis	81
3.4.12	Surgical preparation and chronic miniscope imaging	82
3.4.13	Miniscope Processing Suite (MPS) pipeline	83
3.4.14	Histology and viral injection confirmation	83
3.4.15	Demand-curve fitting and derived behavioral-economics parameters	84

3.4.16	Cross-session DMS dynamics analysis	84
3.4.17	Statistical analysis	86
B	Supplementary material for Chapter 3	89
	Supplementary Information	89
	Supplementary Note 1: Ownership Guard Uniqueness	89
	Supplementary Note 2: $\sqrt{N}$ Threshold Scaling	89
	Supplementary Note 3: Coverage Remediation	91
	Supplementary Note 4: Computational Complexity	92
	Supplementary Note 5: Synthetic Degradation Model	92
	Supplementary Note 6: CellReg Reimplementation and Validation	93
	Supplementary Note 7: Related tools and why they are not head-to-head comparators	94
	Supplementary Note 8: Full Per-Condition Benchmark Statistics	94
4	The Lateral Habenula Retrodictively, Not Predictively, Reports the Outcome of Adverse Consequences Under Fentanyl Self-Administration	96
4.1	Introduction	96
4.2	Methods	99
4.2.1	Animals and housing	99
4.2.2	Oral Fentanyl Self-Administration	99
4.2.3	Changepoint detection of within-session strategy shifts	101
4.2.4	Ultrasonic Vocalization Recording & Analysis	102
4.2.5	Pose Estimation and Kinematic Analysis	103
4.2.6	Surgical Preparation and Chronic Miniscope Imaging	104
4.2.7	Miniscope Processing Suite (MPS) pipeline	104
4.2.8	Stars2Cells: Cross-session registration with Stars2Cells	106
4.2.9	Calcium Signal Quantification and Trial-Aligned Analysis	106
4.2.10	Population decoding of task events and adjacent-trial choice	108
4.2.11	Histology	109
4.2.12	Statistical Analyses	110
4.3	Results	111
4.3.1	Oral fentanyl self-administration produces stable baseline intake independent of sex	111
4.3.2	Rats suppress fentanyl consumption abruptly at high shock intensity	112
4.3.3	Shock shifts the timing of a within-session change in strategy	115
4.3.4	Ultrasonic vocalizations reveal complex affective dynamics during conflicting decision-making	117
4.3.5	Chronic miniscope imaging and cross-session registration of LHb neurons	120
4.3.6	LHb ensembles shift responsiveness and reorganize as shock intensity rises	122
4.3.7	The LHb is not predictive, but retrodictive to a Danger press	124

4.4	Discussion	124
4.4.1	Conventional shock intensities bracket the behavioral transition	126
4.4.2	Ultrasonic vocalizations dissociate affective state from behavioral output . . .	127
4.4.3	Shock modulates the timing of a within-session policy shift	127
4.4.4	Multi-hour recording and cross-session registration require dedicated compu- tational tools	128
4.4.5	LHb ensemble reorganization coincides with the behavioral threshold	129
4.4.6	LHb activity is retrodictive rather than predictive of Danger presses	129
4.4.7	Limitations	130
4.4.8	Implications for decision-making under active drug exposure	131
A	Supplementary material for Chapter 4	133
5	General Discussion	138
5.1	Summary of findings	138
5.2	The Miniscope Processing Suite removes the scale limit	138
5.3	Stars2Cells recovers same-cell identity across sessions	139
5.4	The lateral habenula runs behind the choice	139
5.5	Limitations across the dissertation	142
5.6	Future directions	143
5.7	Conclusion	144
	Bibliography	145

List of Figures

2.1	Overview of the Miniscope Processing Suite (MPS): a no-code, end-to-end pipeline for calcium imaging analysis	12
2.2	Video preprocessing demonstrates effective background subtraction, motion correction, and quality control	14
2.3	Automated field-of-view cropping improves computational efficiency by reducing pixel load	16
2.4	Watershed segmentation and component initialization with spatial merging and validation	17
2.5	Temporal activity update through autoregressive deconvolution with quality control filtering	19
2.6	Spatial footprint refinement through localized nonnegative regression with neighbor-aware updates	21
2.7	Final temporal refinement with updated spatial footprints produces high-quality curated neural components	23
2.8	MPS recovers synthetic ground-truth sources with high spatial and temporal fidelity across recording durations	24
3.1	The longitudinal tracking gap, existing tool landscape, and the GUI-driven S2C application	52
3.2	The Stars2Cells pipeline	57
3.3	Temporal correlation alone is chance-level for cross-session matching	59
3.4	S2C is robust across the full benchmark parameter space and outperforms ROI-based matching at every condition	62
3.5	End-to-end application of S2C via same-cell longitudinal tracking of DMS neurons across an escalating fentanyl demand curve	64
3.6	A conserved DMS population response masks near-complete single-neuron turnover across a fentanyl demand curve	67
4.1	A delayed-shock paradigm for fentanyl self-administration	111
4.2	Rats suppress fentanyl consumption abruptly at high shock intensity	114
4.3	Shock shifts the timing of a within-session change in strategy	116

4.4	Ultrasonic vocalizations reveal complex affective dynamics during conflicting decision-making	119
4.5	Chronic miniscope imaging and cross-session registration of LHb neurons	121
4.6	LHb ensembles shift responsiveness and reorganize as shock intensity rises	123
4.7	The LHb is not predictive, but retrodictive to a Danger press	125
A.1	Per-session distributions and phase-resolved press rates underlying shock-induced suppression	134
A.2	Unsupervised classification of ultrasonic call types and their redistribution across shock intensities	135
A.3	Segmentation yield, per-session count stability, and GRIN-lens targeting across the imaged cohort	136
A.4	No LHb ensembles emerge during the rewarded Safe trials or the No Shock session .	137

List of Tables

2.1	MPS algorithmic innovations and empirical performance across 30 recording sessions. Sessions averaged 95,467 frames (86,713–102,874) at 10 Hz, corresponding to 2.4–2.9 hours of recording. Hardware: Intel [®] Xeon [®] w7-3455, 8 workers with 2 threads each, Dask memory capped at 200 GB.	10
2.2	Variables Produced by Each Step of the Miniscope Processing Suite (MPS).	30
B.1	Per-condition S2C vs. ROI-based matching (worst-case footprint scenario)	95

Chapter 1

General Introduction

1.1 The opioid crisis and decisions made while the drug is on board

The opioid crisis continues to disrupt communities worldwide, with over 100,000 overdose deaths in the United States in 2023 (Garnett and Miniño, 2024) largely attributable to illegally manufactured fentanyl (Volkow and Blanco, 2021). To combat this, a tremendous amount of effort has been put in to study the abstinence model of various substance use disorders where one transitions back from the point of no drug use to drug use. Since most of what we know about the underlying neuroscience comes from these models, the therapies and interventions we build are aimed at the drug-free state rather than at the state in which the drug is already on board (Pickens et al., 2011; Venniro et al., 2016; Marchant et al., 2013; Fredriksson et al., 2021c). Though it is difficult to quantify the exact value, the more common path to overdose stems from continuous use despite adverse consequences and not from sporadic use (Gomez et al., 2026). In this work, continued use is defined as a series of events that unfold in an unwanted way while the drug is already on board (McNally et al., 2023). Thus, if we focus on a framework organized around abstinence, then we cannot discover whether a mechanism for stopping mid episode can be learned at all, since it never puts the animal or person in the state where that mechanism would be used.

This dissertation asks if this is learnable at all and what the lateral habenula (LHb) does at that moment. Therefore, to address this question, we built a delayed shock paradigm for oral fentanyl self-administration where after a period in which the animal has to load to a concentration comparable to its own baseline (Safe Period), any further self-administration produces a 1 s shock that arrives 4 s after each rewarded press (Danger Period), with shock intensity escalated across days from 0.1, 0.2, 0.4, to 0.8 mA in the same animals before a Relapse one week later. By placing the adverse consequence inside that Danger Period, several obstacles were created that needed to be overcome in ways a drug-free procedure does not. Additionally, this paradigm of fentanyl self-administration is structured within a session, as evidenced by animals quickly loading (loading period) to a preferred brain concentration and then responding more slowly to maintain that concentration (maintenance period) (Coffey et al., 2023). Thus, placing the adverse consequence after loading also allows us to observe the animal's change in that strategy while the drug is on

board. However, we needed both to record long enough to hold the drug on board and then register the same neurons held across weeks, and neither capability was available. Chapter 2 builds the first, Chapter 3 builds the second, and Chapter 4 reports that the LHb carries the shock while fentanyl is on board, holds it a week later, and reports the press only after the press has happened.

1.2 The lateral habenula as an outcome signal

A novel aspect of this design was that the consequence has to arrive while the drug is pharmacologically active, which requires a delay between the press and the shock rather than the immediate contingency the operant tradition uses (Azrin and Holz, 1966), which allows us to parse out the two different, but unified, functions of the LHb as a drug-axis signal and a shock axis signal. To explain, the LHb is understood as the canonical negative value hub of the mammalian brain, where neurons fire when an outcome is worse than expected and go quiet when an outcome is better than expected (Matsumoto and Hikosaka, 2007, 2009), and that signal is relayed through the rostromedial tegmental nucleus onto midbrain dopamine neurons and onto serotonergic raphe nuclei (Jhou et al., 2009; Hong et al., 2011; Proulx et al., 2014). Two largely separate literatures have asked what it does when a drug is involved and they have not yet met. On the drug axis, LHb activity tracks the value of the drug itself (Coffey et al., 2023). On the shock axis, LHb neurons respond to externally imposed aversive stimuli (Matsumoto and Hikosaka, 2009; Wang et al., 2017; Congiu et al., 2019; Shabel et al., 2019), inactivation leaves the suppression of shocked instrumental responding intact (Jean-Richard-dit Bressel and McNally, 2014), and habenular engagement covaries with whether an animal keeps responding once its self-administration is shocked (Duan et al., 2022). Neither arrangement reproduces the state that defines opioid use disorder, in which the choice to continue is made while the drug is pharmacologically active and while the consequence is already known. Chapter 4 develops that argument in full.

Notably, the two literatures also diverge on the shock level. Shocked self-administration converged on 0.2 to 0.5 mA (Engeln and Ahmed, 2024; Deroche-Gamonet et al., 2004; Pelloux et al., 2007; Jonkman et al., 2012; Marchant et al., 2018), whereas stress and fear work converged on 0.8 mA by way of the learned helplessness protocol (Vollmayr and Henn, 2001; Li et al., 2011; Yang et al., 2018). Thus, we wanted to design a behavioral paradigm that addressed this gap. Additionally, most *in vivo* work in this structure has used fiber photometry and a consequence of that is that we are assuming that the LHb is responding uniformly to a given stimulus or response. For example, a weak averaged trace is equally consistent with a population in which every neuron responds weakly and with two subpopulations that respond strongly in opposite directions, and an average cannot answer whether the neuron carrying a signal this week carried it last week (Legaria et al., 2022; Engelhard et al., 2019). This uniform or global response seems unlikely given that the LHb's afferents arrive from segregated sources carrying cognitive control, incentive value, homeostatic state, and nociception (Hu et al., 2020; Hikosaka, 2010; Barker et al., 2017), its efferents are split between dopaminergic and serotonergic targets (Hikosaka et al., 2008), and multiple molecular

subtypes coexist within the nucleus (Hashikawa et al., 2020; Wallace et al., 2020). Therefore, by using miniscopes, we were able to both capture neural activity in freely moving animals and access the potential heterogeneous response of a key brain area associated with adverse consequences and reward. Additionally, because the behavioral read-out of an adverse consequence is whether the animal stops, the affective cost of continuing is otherwise inferred from the stopping itself. Ultrasonic vocalizations index affective state independently of responding (Brudzynski, 2013; Coffey et al., 2019), so recording them alongside self-administration lets that cost be measured rather than read off the behavior it is meant to explain.

1.3 The methodological bottleneck of scale and identity across sessions

Miniscopes allow chronic calcium imaging of hundreds of neurons in freely behaving animals (Ghosh et al., 2011a; Cai et al., 2016b; Aharoni and Hoogland, 2019a) and reach deep structures such as the LHb that head-fixed objectives cannot (Ziv and Ghosh, 2015a; Zhou et al., 2018a). However, publication using Miniscopes has remained low due to technical or expertise hurdles. For example, existing pipelines were built for short 15-30 minute recordings and also require a deep computational/engineering background to use (Giovannucci et al., 2019b; Zhou et al., 2018a; Lu et al., 2018; Dong et al., 2022a). Thus, the first technical issue of this dissertation's question is to be able to record long enough to capture the drug onboarding period and what happens throughout the repeated presentation of adverse consequences. Chapter 2 presents the Miniscope Processing Suite (MPS) that removes both limits and solves these issues. Since no ground truth exists, MPS was validated on a synthetic dataset where the ground truth was known and then tested on our recorded data. Moreover, since all existing software fail at the recording length we used, no fair performance comparison can be made.

The second technical issue of this question requires being able to track the same neurons over time, which is essential to understand the consistency, or lack thereof, of a brain area's response. This is important since any treatments or interventions that target that brain area might be ineffective if one brain area is doing two diametrically opposed responses. The mainstream approach to "register" neurons across times uses region-of-interest (ROI) matching based on correlates of per-session footprints, centroid distances, or signal correlations (Sheintuch et al., 2017b; Johnston et al., 2022; Majnik et al., 2025; Vergara et al., 2025a; Hattori and Komiyama, 2022; Lagache et al., 2021), which break under deduction, the realistic dynamics of recording, are not applicable to common recording setups, and/or assumptions about the consistency of neuronal responses that remain unfounded. For instance, in miniscopes, this is difficult since the field of view shifts between mountings and the scope may be remounted in a different orientation. Moreover, this applies outside of only miniscopes as evidenced by the fact that if two neurons are 1px away from each other one recording session, and a new neuron starts firing during the next session 0.5 px away, every single current approach would erroneously mark this new neuron as a match to the previous session. Chapter 3 presents

Stars2Cells (S2C) to address this problem by adapting astrometric plate solving (Hogg et al., 2008; Lang et al., 2010) to recover correspondence from centroid geometry alone. Both tools (MPS & S2C) were released as open-source, no-code, and free tools. Additionally, since no ground truth exists, S2C was also validated on a synthetic dataset where the ground truth was known and then tested on our recorded data and public recorded optical data from the Allen Institute. Moreover, a deep comparison of the existing softwares was able to be made and can be found in this chapter.

To validate the accuracy of S2C on a real world dataset, we recorded from the dorsomedial striatum (DMS) during a standard behavioral economics task during fentanyl self-administration. To our knowledge every published *in vivo* recording in DMS cell bodies during opioid self-administration has been bulk (Seiler et al., 2022; Legaria et al., 2022; Coffey et al., 2023; Bender et al., 2024), and how DMS output neurons reorganize across chronic opioid exposure has been characterized *in vivo* only in nucleus accumbens (Graziane et al., 2016; Hearing et al., 2016; Engeln et al., 2020; Domingues et al., 2025). Representational drift, in which a stable population code is carried by a turning-over set of cells, cannot be distinguished in a bulk signal since it sums across whichever cells are responsive on the day. Moreover, given that: 1) Representational drift has been established in the hippocampus, cortex, and piriform (Ziv et al., 2013a; Driscoll et al., 2017; Schoonover et al., 2021; Rule et al., 2019), 2) The DMS is the canonical associative striatal hub for action-outcome learning, value computation, and goal-directed control (Yin and Knowlton, 2006; Balleine et al., 2021), and 3) The DMS is built from two intermingled projection classes, the direct pathway D1 and indirect pathway D2 medium spiny neurons (Gerfen and Surmeier, 2011; Kreitzer and Malenka, 2008), then the DMS was a strong place in which to ask whether a stable population response is carried by a stable set of cells or by a set that turns over across sessions. Additionally, asking whether the same DMS neurons carry the press response across a behavioral-economics demand curve (Hursh and Silberberg, 2008; Bentzley et al., 2014; Fragale et al., 2019; McConnell et al., 2021; Coffey et al., 2023) is also therefore impossible without same-cell tracking, which is the capability Chapter 3 provides and then uses.

1.4 Tools, terms, and the theories they carry

This dissertation contains no optogenetic manipulation of the lateral habenula, which has become the default tool for causal claims in neuroscience, but contains a more classical behavioral manipulation. To be clear, this is not the same as containing no manipulation, which would only hold if manipulation were narrowed to mean circuit perturbations that remain clinically irrelevant in deeper brain areas in the substance use population. For example, shock intensity was escalated within animal from 0.1 to 0.8 mA, contingency was switched within session between Safe and Danger trials, and the time off the drug was manipulated by a week of forced abstinence before the relapse test in Chapter 4. This distinction matters because it should be the question that determines the tool rather than the reverse. For instance, a brain area that lies downstream of decision making would be theoretically insensitive to manipulations meant to change the decision. Ergo, since this dissertation asks whether LHB

activity runs ahead of, coincides with, or lags behind the decision to press under shock, and found that it ran behind it, then a silencing or driving experiment is rendered uninformative. Moreover, the manipulations available in a person are behavioral ones given that silencing the lateral habenula in a clinical population may be decades away from a version of that intervention that is selective enough to do anything meaningful, assuming all genetic targets or areas respond homogeneously. It is important to highlight that our field often reads circuit perturbation as though behavior were a simple read-out of a circuit when it may be instead the joint output of a circuit and a task the experimenter designed, and that a behavioral result at its own level of description constrains a mechanism rather than standing in for one (Krakauer et al., 2017). If we only ever recorded from the visual cortex while an animal watched music videos, then we would conclude that the visual cortex responds to music, when it is really firing to the images that each song is paired with. This dissertation argues that circuitry recordings are valuable, but often insufficient to glean the whole picture, and thus argues that the field should return to more applicable and useful behavioral manipulation than technically elaborate but insufficient optogenetic manipulation.

The vocabulary this dissertation declines to use is not a stylistic preference. Words like maladaptive, punishment, compulsive, drug seeking, and resistant carry the theory they are meant to test, and the paradigm that supplies the word usually supplies the evidence for it as well. Punishment is defined in the operant tradition by its effect on the response, so a shock is a punisher only if responding decreases (Azrin and Holz, 1966). At 0.4 mA responding here was indistinguishable from baseline, which leaves the word either contradicted or redefined, and neither is an adequate description of the outcome. The terms used throughout are shock, danger, adverse consequence, and self-administration, each of which names something the apparatus did rather than something the word’s theory says the animal became. For instance, a method calibrated to leave a resistant subgroup will find a resistant subgroup. Furthermore, a population that allocates effort toward a reliable high-magnitude reinforcer in a choice set whose alternatives are weaker is doing what optimization under constraints predicts and calling the circuit maladaptive is a category error. Moreover, encodes and represents are also avoided for a related reason. As discussed earlier, a signal that correlates with an experimental parameter under one contingency is not thereby a representation of that parameter, since the code depends on details of the experiment that the spike train does not carry (Brette, 2019; Ebitz and Hayden, 2021). Thus, the decoders in Chapter 4 are therefore described as recovering what is present in the population rather than as reading what the LHb represents, and their ROC-AUC values are reported as a floor on recoverable information rather than as a measurement of content. They also decode a cue, a press, and a shock, which are physical events the apparatus delivered and timestamped, rather than categories such as craving or compulsivity that the experimenter would have had to impute before the classifier could find them (Ritchie et al., 2019).

1.5 Dissertation aims

The first aim was to remove the scale limit on single-photon calcium imaging. Chapter 2 presents MPS, a GUI-driven pipeline that carries raw miniscope video through preprocessing, source extraction, deconvolution, and quality control without code. MPS processed 7.26 TB of data, comprising 77 h of raw video, in 55.6 h on a single workstation capped at 200 GB of memory, and its component recovery was validated against synthetic ground truth.

The second aim was to remove the identity limit and then to spend it on the identity question. Chapter 3 presents S2C, which describes each neuron by the geometry of its four-neuron neighborhood, matches those descriptors under rotation, translation, and uniform scaling, verifies with RANSAC, and assigns correspondences by Hungarian matching. S2C reached $F1 = 98.4\%$ across a 1,262-run synthetic benchmark with construction-time ground truth, reproduced the Allen Brain Observatory’s own registration across 431 two-photon containers at a median $F1$ of 0.97, and retained $41.6 \pm 3.4\%$ of first-session DMS neurons across a four-session chain against $13.0 \pm 3.9\%$ for footprint matching. From this, we found that the press-responsive fraction in the DMS held near one half across a fentanyl demand curve while roughly half of tracked neurons switched class at every transition, so the conserved response in this structure is a property of the group and not of its membership.

The third aim was the biology the tools were built for. Chapter 4 asks whether the LHb signals an adverse consequence during a shift in the strategy of fentanyl self-administration while the drug is still on board, whether that signal runs ahead of the shift or behind it, whether it survives across weeks, and whether the same neurons carry it.

Chapter 2

A No-code and Improved Pipeline for Long-Duration Miniscope Calcium Imaging

Published as: Peden-Asarch A, Weinstock M, Coffey KR, Neumaier JF. Miniscope Processing Suite (MPS): An Intuitive, No-Code, Scalable Pipeline for Long-Duration Calcium Imaging. eLife Reviewed Preprint.

2.1 Introduction

Miniaturized one-photon fluorescence microscopes (“miniscopes”) have revolutionized *in vivo* neural imaging by enabling single-neuron resolution recordings in freely behaving animals (Ghosh et al. (2011b); Ziv et al. (2013b)). Unlike tethered two-photon setups, miniscopes are lightweight enough to be head-mounted on rodents, permitting neural activity to be monitored during natural behaviors without restraining the subject (Aharoni and Hoogland (2019b)). Moreover, through implanted microendoscopic gradient-index (GRIN) lenses, researchers can perform chronic imaging of previously inaccessible cell populations deep in the brain over weeks to months (Ziv and Ghosh (2015b); Zhou et al. (2018b)). As a result, researchers can theoretically track the same neurons across multiple sessions and then study evolving neuronal dynamics (Ziv et al. (2013b); Cai et al. (2016a)). Additionally, recent hardware advances have further enhanced miniscope capabilities with newer designs that support simultaneous dual-wavelength imaging (e.g. recording green and red indicators in parallel) to distinguish cell populations or combine activity reporters with stable reference markers (Barbera et al. (2024); Dong et al. (2025)). Thus, miniscopes integrate single-cell resolution with behavioral compatibility and chronic repeatability and can be implemented in conventional behavioral procedures over repeated days, making them a powerful tool for neuroscience.

The proliferation of miniscope users created a pressing need for user-friendly, scalable analysis pipelines. One-photon calcium imaging data pose special challenges: signals are often confounded

by movement artifacts, background neuropil autofluorescence, and overlapping neuronal footprints, making automated source extraction difficult (Zhou et al. (2018b)). Several computational pipelines have emerged to address these issues, each offering valuable algorithms but with notable practical limitations. To our knowledge, the longest continuous miniscope recordings processed to date span ~ 45 – 60 minutes per session across longitudinal studies extending up to several months (Ziv et al. (2013b); Cai et al. (2016a); Sheintuch et al. (2017a); Gonzalez et al. (2019); Jennings et al. (2019); Chen et al. (2023); Zhou et al. (2020); Guo et al. (2023); Wirtshafter and Disterhoft (2022); Blair et al. (2023); Vergara et al. (2025b)). Theoretically, current pipelines can handle longer behavioral experiments by chunking datasets into separate files, but this can compound errors in motion correction and fails to preserve the true longitudinal stability of neuronal identities over time - a limitation stemming largely from the constraints of existing software. CaImAn is a widely used open-source library that implements motion correction and constrained non-negative matrix factorization (CNMF) for source separation and achieves near-human accuracy in detecting active neurons (Giovannucci et al. (2019a)). However, CaImAn and related packages like CNMF-E (an optimized CNMF for microendoscope data) require users to write code or scripts in MATLAB/Python and adjust parameters programmatically and demand navigating a difficult setup process across Integrated Development Environments (IDEs), virtual environments, and multiple package managers (Giovannucci et al. (2019a); Zhou et al. (2018b)). Other pipelines such as EXTRACT (a GPU-accelerated cell extraction algorithm primarily designed for two-photon datasets, lacking the robust motion correction needed for miniscope recordings) and MIN1PIPE (which introduced morphological background removal and seeded CNMF initialization) likewise demand considerable coding or technical expertise to operate (Lu et al. (2018); Inan et al. (2021b)). More recently, Minian/CaliAli provided an “open-source miniscope analysis pipeline” with interactive visualization of parameters to guide users with minimal computational experience (Dong et al. (2022b); Vergara et al. (2025b)). These softwares reduce memory usage by leveraging Dask’s out-of-core computation, allowing analysis to run on consumer-grade hardware. Nonetheless, this strategy comes with trade-offs: memory is fragmented across processes and not released until the script or IDE is fully closed, and long recordings eventually collapse into a single oversized task graph. For example, a typical 600×600 field of view recorded for 95,000 frames (~ 3 hours) already produces 34.2×10^9 data points. When multiplied across $\sim 1,000$ components for matrix factorization, this balloons to 34.2×10^{12} elements – requiring more than ~ 270 TB of memory if handled naively in 64-bit precision. In practice, this Dask graph attempts to juggle this scale of computation, but since the entire dataset is put into memory at once at a key CNMF related step, the processing of datasets of this size is impossible. Users must also still launch Jupyter notebooks or scripts, rather than relying on a stable standalone GUI. (Dong et al. (2022b)) Additionally, even given a machine with ~ 512 GB of RAM, the largest video size successfully run before saturating memory and causing a crash was a 65 minute video with 600×600 pixels with MIN1PIPE. In practice, running and maintaining pipelines of similar scale often demands deep computer science expertise (managing virtual environments, CUDA dependencies, and package conflicts) that many

biology labs rightfully lack. Lastly, deep-learning-based ROI detection has emerged in recent software; MPS is complementary to these approaches and supports integration of lab-specific models (Tian et al. (2025)). Therefore, MPS directly addresses this gap by integrating state-of-the-art algorithms into a stable, point-and-click platform engineered to process multi-hour recordings on everyday workstations. No existing solution fully combines advanced calcium imaging algorithms with an accessible, graphical user interface (GUI) that can reliably handle large datasets without overwhelming standard laboratory hardware. As many labs find data processing to be a bottleneck and lack the coding expertise to take full advantage of miniscope technology, it is not surprising then to see that miniscopes have been underutilized by the field at large.

In this paper, we introduce the Miniscope Processing Suite (MPS), a GUI-based, end-to-end calcium imaging pipeline optimized for long-duration miniscope recordings. MPS builds on the gold-standard CNMF framework (Pnevmatikakis et al. (2016a)), but adapts it for long-duration miniscope data. Beyond Non-negative Double Singular Value Decomposition (NDSVD) initialization and custom multi-lasso regularization, MPS introduces (i) robust preprocessing with neuropil subtraction, denoising, and cross-correlation motion correction; (ii) interactive regions of interest (ROI) cropping to reduce pixel load by $\sim 56\%$ (median); (iii) watershed-based segmentation with merging/validation to avoid duplicate or fragmented ROIs; (iv) parallelized temporal and spatial CNMF updates with out-of-core Dask execution, enabling stable analysis of multi-hour datasets; (v) automated artifact detection (e.g., erroneous frames and line-splitting events); and (vi) a built-in GUI Data Explorer for final quality control. These combined improvements allow MPS to scale where prior pipelines fail, while remaining accessible to non-programmers. Additionally, The pipeline is built on a Dask parallel computing framework, similar to Minian, enabling scalable processing of large datasets on standard laboratory hardware (e.g. multi-core PCs) without requiring high-performance clusters (Dong et al. (2022b)). Unlike prior pipelines, MPS enables users to complete every step via an intuitive GUI, created in collaboration with experts in UI/UX design. By combining powerful algorithms with an intuitive interface, MPS aims to lower the barrier for miniscope data analysis revolutionizing the field. This paper describes the design of MPS and demonstrates that it can reliably analyze multi-hour recordings with hundreds of neurons, filling an important need for a scalable and accessible miniscope imaging pipeline. We anticipate that MPS will greatly facilitate the adoption of miniscope technology by the broader community and promote more standardized, reproducible analysis practices.

2.2 Results

2.2.1 Pipeline Overview & Installation (Figure 1: Step 1)

We developed the Miniscope Processing Suite (MPS), a No-Code GUI-driven pipeline for end-to-end preprocessing, source extraction, and curation of long-duration one-photon calcium imaging that takes an AVI video file from a one-photon imaging experiment and returns a processed temporal and spatial matrix, relevant background components, and a plethora of other intermediate step

variables (Figure 2.1A). Unlike CaImAn (Giovannucci et al. (2019a)), Minian (Dong et al. (2022b)), or Min1Pipe (Lu et al. (2018)), MPS installs via a standalone launcher and runs entirely through a point-and-click interface (https://ariasarch.github.io/MPS_Installer/). Across datasets up to ~ 2.7 hours ($\sim 96,000$ frames), MPS completed all stages on a single workstation (Intel[®] Xeon[®] w7-3455, Dask memory capped at ~ 200 GB) with a totality of ~ 7.26 TB / ~ 77 hours of video in a mere 55.6 hours to complete from start to finish. This makes MPS one of the first fully GUI-based pipelines capable of processing tens of terabytes of long-duration miniscope data on a single workstation, highlighting both its accessibility and computational power (Table 2.1).

Table 2.1: MPS algorithmic innovations and empirical performance across 30 recording sessions. Sessions averaged 95,467 frames (86,713–102,874) at 10 Hz, corresponding to 2.4–2.9 hours of recording. Hardware: Intel[®] Xeon[®] w7-3455, 8 workers with 2 threads each, Dask memory capped at 200 GB.

Pipeline Stage	Traditional Approach	MPS Approach	Median Runtime (min)	% Total
Preprocessing	FFT-based motion correction	+ line-split detection + erroneous frame removal	37 (IQR: 36–40)	32%
Initialization	$2 \times$ least-squares on flattened data [$2 \times M \times T$ in RAM]	NNDSVD + watershed + correlation merging [$0.3M \times T$]	26 (IQR: 21–39)	23%
Cell Detection	Direct segmentation	Multi-scale watershed + spatial/temporal merging	13 (IQR: 8–26)	11%
Validation	Basic filtering	Size/shape/overlap + trace validation + noise estimation	4 (IQR: 3–6)	3%
CNMF: Temporal	Spatial-first (update all A , then all C)	Temporal-first (filter early for fewer components for spatial)	7 (IQR: 6–8)	6%
CNMF: Spatial	Global updates on full FOV [$K \times M$ problem]	KD-tree windowing [$K_{\text{local}} \times 5\text{--}20\%$ pixels]	7 (IQR: 5–16)	6%
Export	Standard file I/O	Chunked Zarr, JSON, Pickle, and NPY file export	7 (IQR: 5–8)	6%
TOTAL	Sequential LS-LS-CNMF architecture	Parallel NNDSVD-temporal-spatial architecture	111 (IQR: 96–137)	100%

Processing rate: MPS achieves 0.7–0.9 min/min of recording. Across 30 sessions totaling ~ 7.26 TB / ~ 77 hours of video (up to 2.9 hours per session), MPS completed all stages in 55.6 hours, making it one of the first fully GUI-based pipelines capable of processing tens of terabytes of long-duration miniscope data on a single workstation. Traditional approaches are not directly comparable on multi-hour recordings, as they fail to complete on datasets of this duration due to memory exhaustion or graph bloat. MPS achieves 0.7–0.9 min/min on 3-hour sessions where no reference output can be produced by existing pipelines.

MPS presents a stepwise GUI (Steps 1-8) with progress bars and easily digestible logs (Figure 2.1B). Purple modules reflect standard preprocessing steps used across existing pipelines: Deglow, Denoise (x2), and Noise Estimation. Green modules represent steps introduced in MPS, including Window Cropping, NNDSVD-seeded initialization, Watershed segmentation, Spatial Merge, Signal Extraction, and Temporal Merge. NNDSVD is a standard nonnegative initialization method, used here to seed CNMF directly in place of least-squares or greedy ROI-based initialization. Blue modules are heavily re-engineered versions of existing algorithms, rewritten to scale to three-hour videos via parallelization: Motion Correction with Frame Removal and iterative CNMF (Spatial/Temporal) updates. Together, this workflow enables MPS to deliver a complete, GUI-driven pipeline for

long-duration miniscope imaging. Across 25 installs on macOS and Windows, median setup was 2.5 minutes, and all installs completed once the payload was written to a user-owned directory. Installs targeting protected system directories on macOS can fail with an `rsync` permission error, so the installer now validates the target path before unpacking.

The Miniscope Data Explorer provides an intuitive environment for quality control after loading spatial footprints (A) and calcium traces (C). Users can interactively inspect each component, visualize its spatial map alongside its temporal activity, and remove low-quality or artifactual cells directly within the interface (Figure 2.1C). A synchronized movie player allows firing activity to be overlaid in real time, enabling visual confirmation of trace fidelity without the need to reconfigure or rerun earlier analysis steps. All edits are non-destructive and can be saved out at any point. The system automatically writes 93 structured files per session, ensuring that data can be re-analyzed globally or revisited on a per-component basis. This design provides both flexibility and reproducibility: investigators can iteratively refine their dataset while maintaining the ability to restore or compare with prior states (Table 2.2).

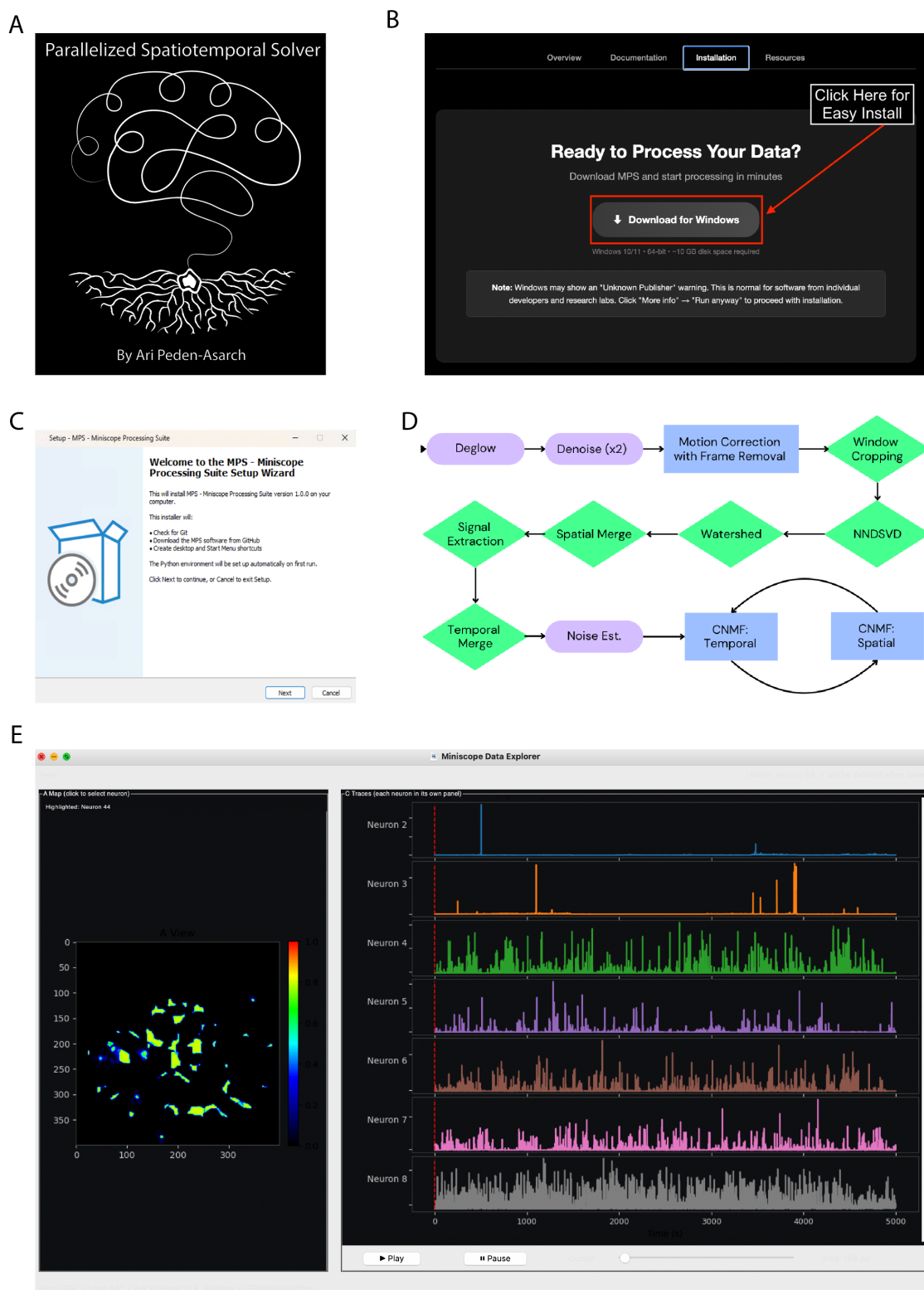


Figure 2.1: Overview of the Miniscope Processing Suite (MPS): a no-code, end-to-end pipeline for calcium imaging analysis. **(A)** Logo drawn by Ari Peden-Asarch. **(B)** Processing pipeline flowchart illustrating the modular architecture of MPS. **(C)** Interactive data explorer for visualization and quality control of extracted neural signals. Left panel displays spatial footprints of identified neurons color-coded by activity level, with pixel coordinate axes and heatmap scale showing fluorescence intensity. Right panel shows temporal activity traces for selected neurons over 5000 frames.

2.2.2 Video Preprocessing (Figure 2: Steps 2a-2f)

The preprocessing pipeline standardizes video input (Figure 2.2A), subtracts neuropil/background and denoises (Figure 2.2B), performs rigid cross-correlation motion correction (Figure 2.2C), and excludes erroneous frames (line-splits, blur, saturation). On 30 recordings (FOV 600×600 pixels, 10 Hz), background suppression reduced the median baseline by 92.59%, and cell-to-surround contrast increased 0.033x (Figure 2.2F). This large reduction is important because neuropil fluorescence dominates raw one-photon miniscope signals; without suppression, somatic activity can be masked by widespread background fluctuations, leading to artificially inflated correlations and reduced sensitivity to single-cell events. Removing this background markedly improves the detectability and interpretability of calcium transients, in line with prior work emphasizing neuropil subtraction in miniscope analysis. Motion correction yielded a residual shift of 3.081 pixels (IQR 2.074 - 4.001) and a long-term drift of ~ 0.571 pixels/hr (Figure 2.2G-I). Correcting for motion is critical, as even sub-pixel jitter accumulates over long recordings and can blur spatial footprints, lowering signal-to-noise ratio. Residual drift across hours is especially problematic for neuron registration and stability analysis; mitigating this motion ensures that cell footprints remain well-aligned and neural activity can be tracked reliably across the session.

Frame quality control flagged on average 0.37% of frames (Figure 2.2I). Although a small proportion, erroneous frames - such as blank captures when the device fails to shut off cleanly, or distorted frames from commutator strain - can strongly skew long recordings (Figure 2.2E). Over three-hour sessions, even rare outliers can bias mean fluorescence, corrupt correlation structure, or disrupt motion correction templates. Automated detection and removal of these frames prevents such disproportionate effects and helps stabilize downstream analyses. Line splitting artifacts are where blocks of rows within a frame are circularly shifted or stitched from an adjacent temporal frame due to transient disruptions in the high-speed serial data link during sequential row readout (a data-transport corruption, not rolling-shutter motion). MPS detects these artifacts at the frame level from the intensity signature they leave at the affected rows or from frames whose left-edge intensity (or, optionally, whose detrended row-profile residual) exceeds a robust threshold and are automatically flagged and removed (Figure 2.2D). We detect and exclude these frames rather than attempt reconstruction (see Methods, Step 2a).

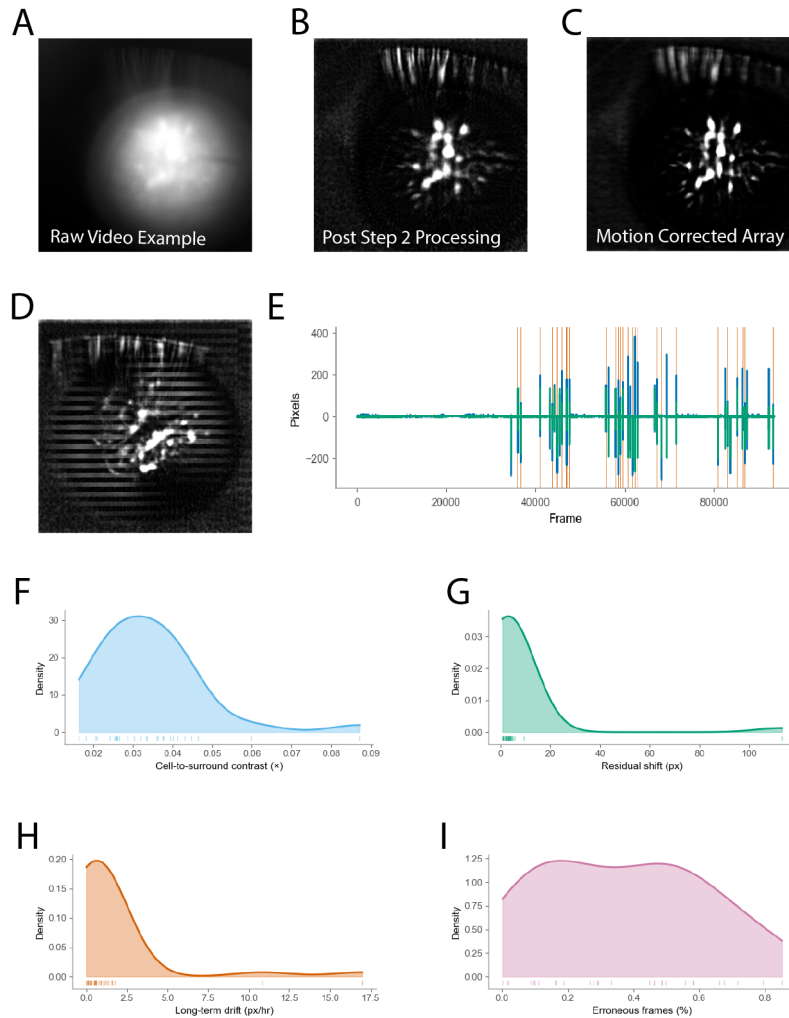


Figure 2.2: Video preprocessing demonstrates effective background subtraction, motion correction, and quality control. **(A)** Raw miniscope frame showing diffuse background fluorescence and vignetting that obscures individual cells. **(B)** Same frame after background subtraction and denoising, revealing individual neuronal somata with improved contrast. **(C)** Motion-corrected frame showing stable spatial alignment of cellular structures. **(D)** Concatenated view across multiple frames illustrating line-splitting artifacts (horizontal discontinuities) that are automatically detected and removed. **(E)** Per-frame motion estimates in X (blue) and Y (yellow) dimensions across the full recording, with erroneous frames highlighted in red bars indicating frames flagged for excessive motion or artifacts ($n=518$ frames, 0.59% of total of this session). **(F)** Distribution of cell-to-surround contrast showing improvement after preprocessing (median increase of $0.033\times$), with individual data points shown as rug plot. **(G)** Distribution of residual shift after motion correction (median = 3.081 pixels, IQR 2.074–4.001 pixels). **(H)** Long-term drift quantified as displacement over time (median = 0.571 pixels/hour), demonstrating stability of motion correction across extended recordings. **(I)** Distribution of erroneous frames as percentage of total recording across all sessions (median = 0.37%), showing automated quality control effectively identifies and removes artifacts without substantial data loss.

2.2.3 ROI Cropping & Dimensionality Reduction (Figure 3: Steps 3a-3c)

Interactive cropping is applied to remove peripheral or dark regions (Figure 2.3A), which cut the pixel load by 56% (median, $59.0 \pm 6.6\%$ mean $\pm$ SD across 30 sessions). This reduction

decreased runtime by 2.2x (median, 2.52 ± 0.52 x mean $\pm$ SD) and lowered peak memory usage by 56% (median). Cropping is important because dark borders and unused regions contribute no biological signal but substantially inflate computation time and memory requirements; removing them streamlines analysis without affecting data quality. After cropping, truncated NNDSVD was performed on downsampled frames. With $K = 100$ components, the decomposition captured 100% of the variance ($100.0 \pm 0.0\%$ mean $\pm$ SD across 30 sessions), with the elbow point typically at $K = 3$ (range 3-4) (Figure 2.3B-C). Using NNDSVD ensures that the initial spatial and temporal factors are nonnegative and biologically interpretable, while also reducing the number of CNMF iterations compared to random or unconstrained least-squares starts and example of the top three components are displayed for reference (Figure 2.3D). Thus, these reductions enabled full-session analyses to complete 2-3x faster on average while preserving 100% of the explained variance, demonstrating that substantial computational gains can be achieved without any loss in data. Fluorescence photobleaching over the 3-hour recording period was minimal, with an average signal decay of 17.4% (Figure 2.3E)). It is important to note that these videos were roughly 60 seconds with a five second off period between them over the 3 hour session to reduce chances of overheating and reduce photobleaching. This gradual decline is automatically corrected by the analysis software's baseline normalization algorithm, ensuring reliable quantification throughout extended imaging sessions.

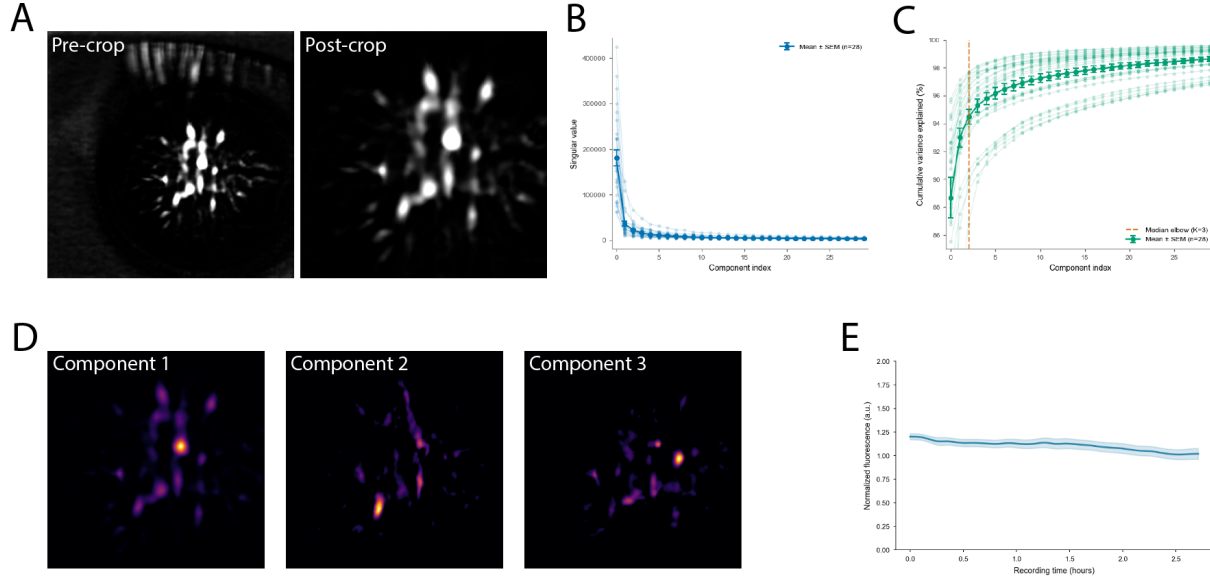


Figure 2.3: Automated field-of-view cropping improves computational efficiency by reducing pixel load. **(A)** Pre-crop miniscope frame showing full field of view with vignetting and peripheral regions lacking cellular activity. **(B)** Post-crop frame demonstrating automated detection and retention of active cellular regions while removing uninformative periphery. **(C)** Singular value spectrum from NNDSVD decomposition showing rapid decay after initial components, with elbow point indicating optimal dimensionality for downstream processing. Green points indicate selected components capturing majority of signal variance. **(D)** First three spatial components from NNDSVD decomposition, revealing distinct cellular structures and spatial patterns that serve as initialization for CNMF. Components are ordered by variance explained, with Component 1 capturing the most prominent spatial features and subsequent components revealing progressively finer structures. **(E)** Fluorescence stability over 3-hour recording period demonstrates minimal photobleaching with mean signal decay of 17.4% with 5s gaps between 60s recordings.

2.2.4 Cell detection, Initialization, and Pre-CNMF Validation (Figure 4: Steps 4a-5b)

Cells are seeded using local peak detection on maximum and percentile projections, followed by watershed segmentation. Over-segmented masks are then merged by spatial overlap (Figure 2.4A). Initial temporal traces are extracted for each mask and combined with NNDSVD factors to form nonnegative, structured starting matrices for spatial and temporal activity. Components with invalid values are removed, and duplicates are merged if they show both high temporal correlation and spatial proximity (Figure 2.4B) & (Figure 2.4D). In Step 5a, MPS calculates a per-pixel noise map by taking the standard deviation of residuals across time after subtracting a low-rank background estimate. The noise map can be optionally rescaled in regions with strong background activity and smoothed with a Gaussian filter for stability. This provides a component-specific noise level that is later used to scale the sparsity penalty during temporal deconvolution (Figure 2.4F–G). Initialization produced 116 ± 39 candidates per field of view (median area 552 pixels) with a duplicate-merge rate of 98.0% (Figure 2.4B). Across sessions, this pipeline reduced the raw watershed output (5827.2 ± 1359.0 small masks, median size 124.3 ± 20.3 pixels) to 116.0 ± 39.8 curated components (median

size 570.7 ± 169.7 pixels) with a 98% duplicate-merge rate, while producing stable noise estimates (overall mean 0.5248 ± 0.0255 ; range 0.3681- 0.9605) (Figure 2.4C) & (Figure 2.4E). Thus, seeding CNMF with NNDSVD-derived nonnegative factors rather than a least squares initialization captures nearly 100% of the variance and reaches that point hours faster.

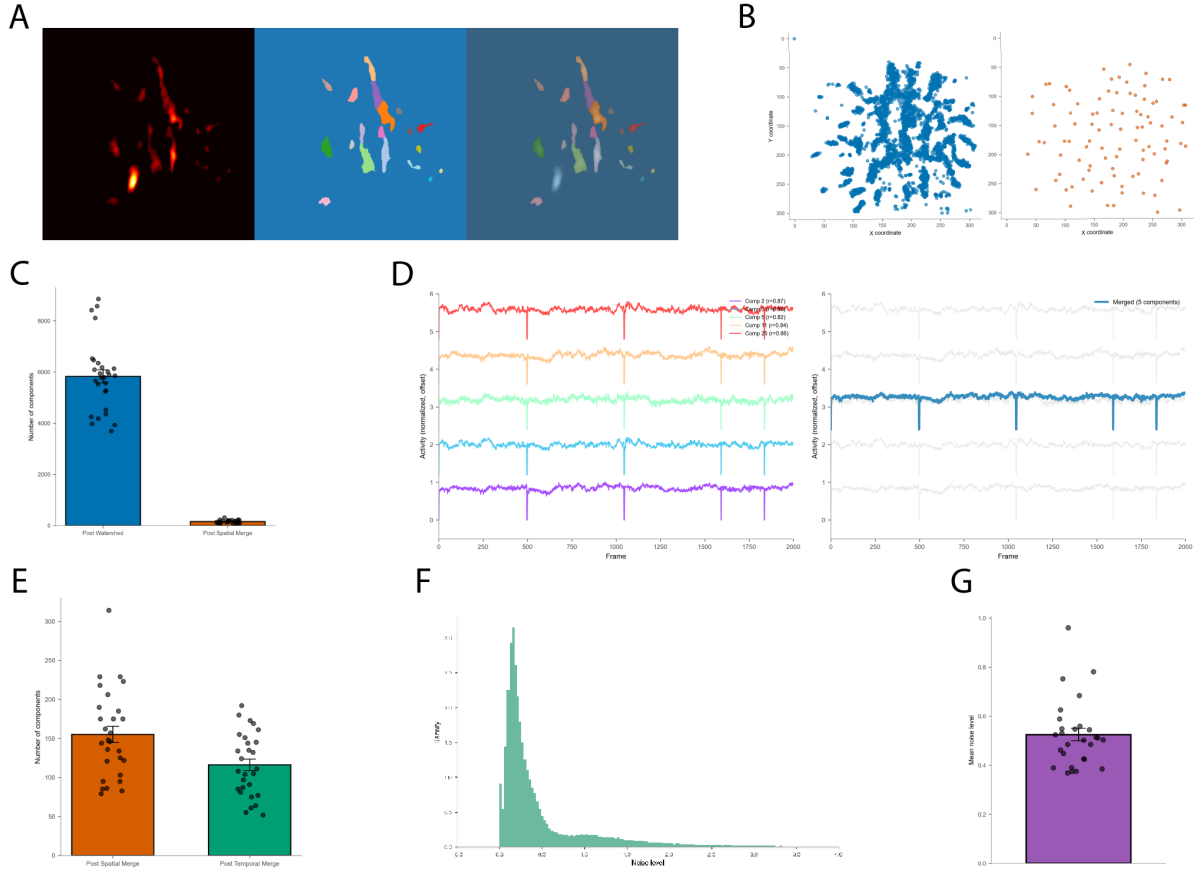


Figure 2.4: Watershed segmentation and component initialization with spatial merging and validation. **(A)** Spatial distribution of detected components for Component 1 of an example session showing raw fluorescence image (left), watershed-segmented components with color-coded masks (middle), and merged overlay (right). **(B)** Scatter plots showing component centroids for all NNDSVD components before (left, high density) and after (right, reduced density) spatial merging. **(C)** Quantification of component counts across the initialization pipeline, showing reduction from pre-watershed (left) to post-spatial merge (right). Bar plot compares before versus after merging across all sessions with individual data points overlaid. **(D)** Representative temporal traces from five example components (Comp 2 through Comp 9) showing calcium activity patterns over 2000 frames. Left panel displays raw temporal traces with distinct activity patterns across components; right panel shows activity trace for merged components, demonstrating successful consolidation of correlated signals. **(E)** Comparison of component counts post-spatial merge (left median = 155.3 ± 53.8) versus post-temporal merge (right, median = 115.3 ± 38.7), demonstrating temporal correlation-based merging further reduces redundant components. **(F)** Distribution of noise level across all sessions, showing strong peak near real signal and an additional peak towards zero, indicating noise remaining that MPS caught. **(G)** Distribution of noise estimates caught across all final initialized components after validation, showing median noise level of approximately 0.52.

2.2.5 Temporal update (Figure 5: Steps 6a-6e)

Temporal activity was updated by projecting the spatial footprints back onto the movie in overlapping time chunks and solving a constrained autoregressive deconvolution. The optimization jointly estimated calcium activity, a nonnegative innovation term with an L1 penalty scaled by each component’s noise level, and baseline terms, subject to nonnegativity and the autoregressive constraint. MPS also suggested decay and sparsity parameters from the noise map and filtered out low-quality components during this step. The residual term (YrA) was centered around zero, indicating that demixing and background subtraction had successfully removed global structure and preserved only noise (Figure 2.5A). On average, components decreased from 114.8 ± 7.6 before filtering to 104.7 ± 6.7 after filtering, corresponding to an overall rejection of 8.8 percent (Figure 2.5E). Median decay time was 1.0 frames with a mean of 1.6 frames, ranging from 0.4 to 20.3 frames across sessions. Signal quality improved markedly, with a mean increase of 55.6 ± 0.8 decibels in signal-to-noise ratio, where signal power was taken as the variance of the trace and noise power was estimated from the standard deviation of the trace derivative (Figure 2.5D). Autoregressive decay coefficients yielded median decay of 1.0 frames (100 ms at 10 Hz) and mean of 1.6 frames (160 ms), consistent with the fast kinetics of GCaMP8f indicators (Zhang et al. (2023b)). Decay times were derived from the autoregressive parameters and summarized with mean, median, and range. It is important to mention that this first pass still yields some noise artifacts, and thus, the subsequent passes described below are needed (Figure 2.5B–C).

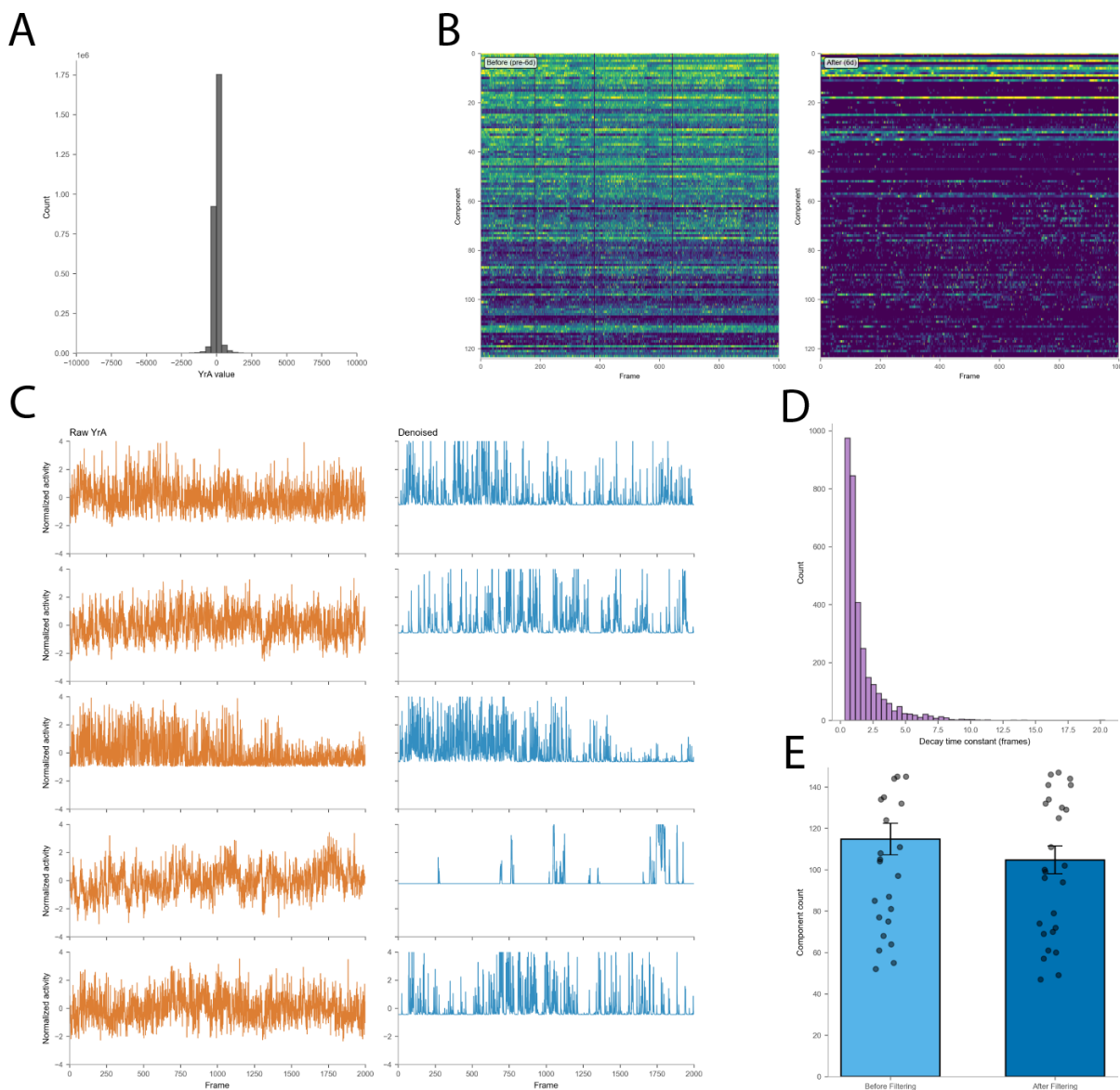


Figure 2.5: Temporal activity update through autoregressive deconvolution with quality control filtering. **(A)** Distribution of YrA values (residual after projecting spatial components onto the movie) centered near zero, indicating successful demixing and background subtraction. **(B)** Heatmaps of temporal activity for all components for a single session before (left) and after (right) the temporal update. Each row is one component. Color intensity represents normalized calcium activity across frames. **(C)** Representative example traces showing the initialized trace (left, 5 examples) and trace after the temporal update (right, 5 examples labeled). **(D)** Distribution of estimated decay time constants across all components, showing median decay of 1.0 frames (mean = 1.6 frames, range 0.4-20.3 frames). **(E)** Component retention across temporal filtering, comparing counts before (left, median = 114.8 ± 7.6) versus after (right, median = 104.7 ± 6.7) quality control, representing 8.8% rejection rate. Signal-to-noise ratio improved by mean of 55.6 ± 0.8 dB across retained components.

2.2.6 Spatial update (Figure 6: Steps 7a-7f)

Spatial footprints were refined using localized, neighbor-aware nonnegative regression. Each component's region was dilated, local residuals were recomputed with neighbor subtraction, and footprints were updated under nonnegativity and locality constraints (Figure 2.6A). Heuristic checks flagged components with implausible sizes, updated masks were adopted, and spatial duplicates were merged. Component counts dropped substantially, from 104.7 ± 6.7 before refinement to 51.6 ± 4.1 after, a reduction of 50.8 percent (Figure 2.6B). Nearest-neighbor statistics showed increased spacing, with mean centroid distance rising from 19.6 ± 10.5 to 27.0 ± 14.5 pixels and median distance increasing from 19.0 to 26.2 pixels (Figure 2.6C). After a single update on the spatial dimension, the location of the cells can now easily be discerned.

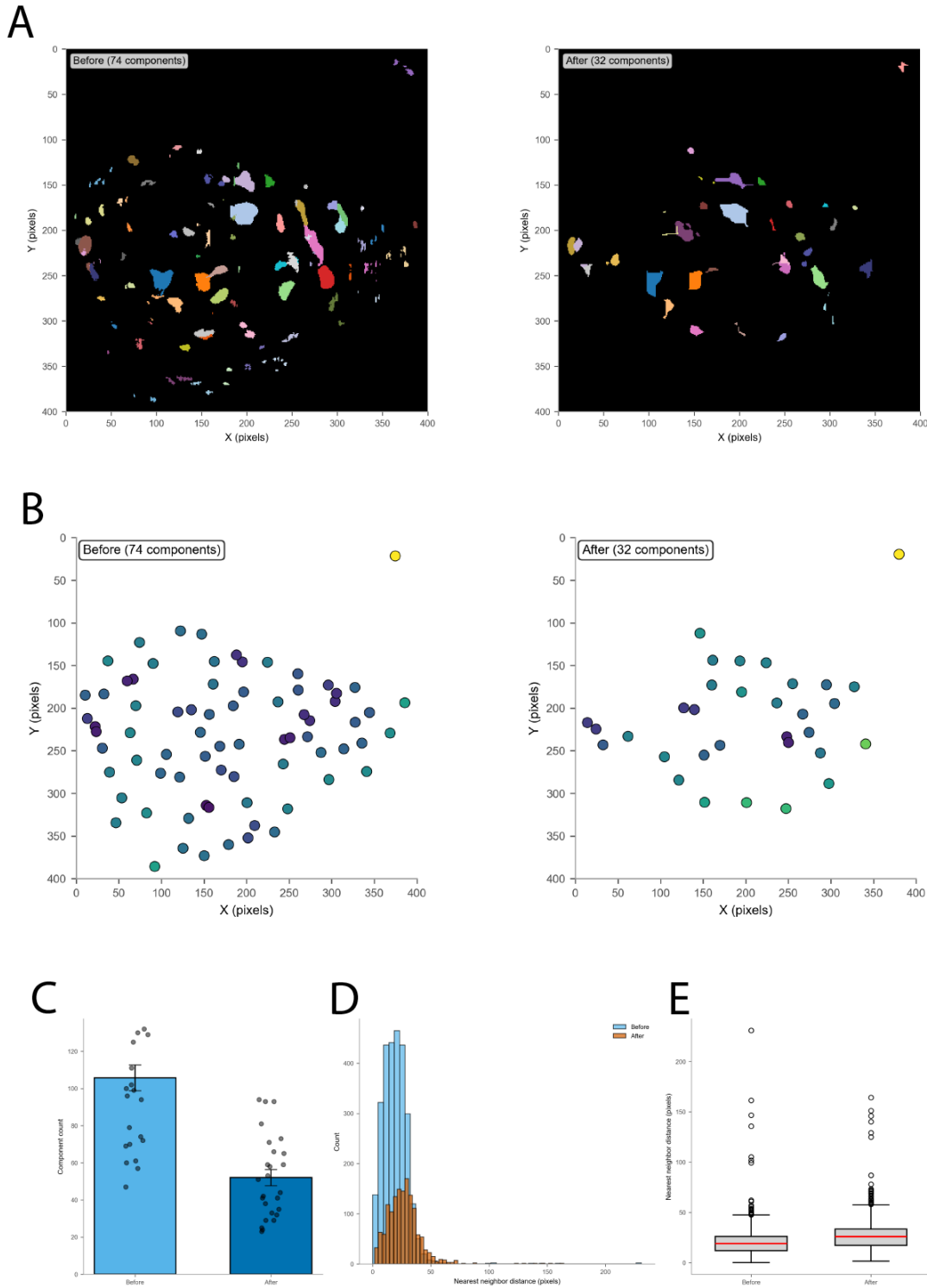


Figure 2.6: Spatial footprint refinement through localized nonnegative regression with neighbor-aware updates. **(A)** Spatial distribution of components before (left) and after (right) spatial refinement. Each colored region represents refined spatial footprint after dilation-bounded regression and neighbor subtraction. Post-refinement components show improved spatial localization and reduced overlap. **(B)** Component count reduction across sessions from pre-spatial update (median = 104.7 ± 6.7) to post-spatial update (median = 51.6 ± 4.1), representing 50.8% reduction through merging of spatial duplicates and removal of implausible footprints. **(C)** Distribution of nearest-neighbor distances between refined component centroids, showing shift toward larger separations with peak around 100-150 pixels and histogram counts up to 400 components.

2.2.7 Final Temporal Update (Figure 7: Steps 8a-8c)

The final temporal refinement was performed after spatial components had been merged and residuals refreshed. This two-stage temporal refinement strategy (updating traces both before and after spatial refinement) follows the iterative optimization protocol established in CNMF-E, where one additional matrix update is typically performed following component interventions to ensure convergence (Zhou et al. (2018b); Giovannucci et al. (2019a)). As in the earlier update, the spatial footprints were projected back onto the movie in overlapping time chunks, and a constrained autoregressive deconvolution was solved to jointly estimate calcium activity, a nonnegative innovation term with sparsity regularization, and baseline parameters. At this stage, the optimization incorporated the merged spatial set and the updated residuals, providing a definitive estimate of temporal activity. Low-quality traces were pruned, and background terms could be optionally modeled alongside component activity. The residual term (YrA) was even more centered around zero, indicating that demixing and background subtraction had better removed global structure and preserved only noise when compared to the first temporal pass (Figure 2.7A) This final pass substantially reduced the number of surviving components, from 117.1 ± 7.7 before filtering to 52.6 ± 4.2 after filtering, corresponding to a 55.1% rejection rate (Figure 2.7E). Median decay time remained 1.0 frames with a mean of 1.6 frames, spanning 0.4–20.3 frames across sessions (Figure 2.7D). Signal quality was further improved, with a mean increase of 47.8 ± 1.2 dB in signal-to-noise ratio. These results established the final set of temporal traces used in downstream analyses (Figure 2.7B–C). Additionally, if one wishes to continue to perform temporal or spatial updates, they are more than able to through MPS, but not required. See Supplementary Video 1 for an example of the final Miniscope Processing Suite output.

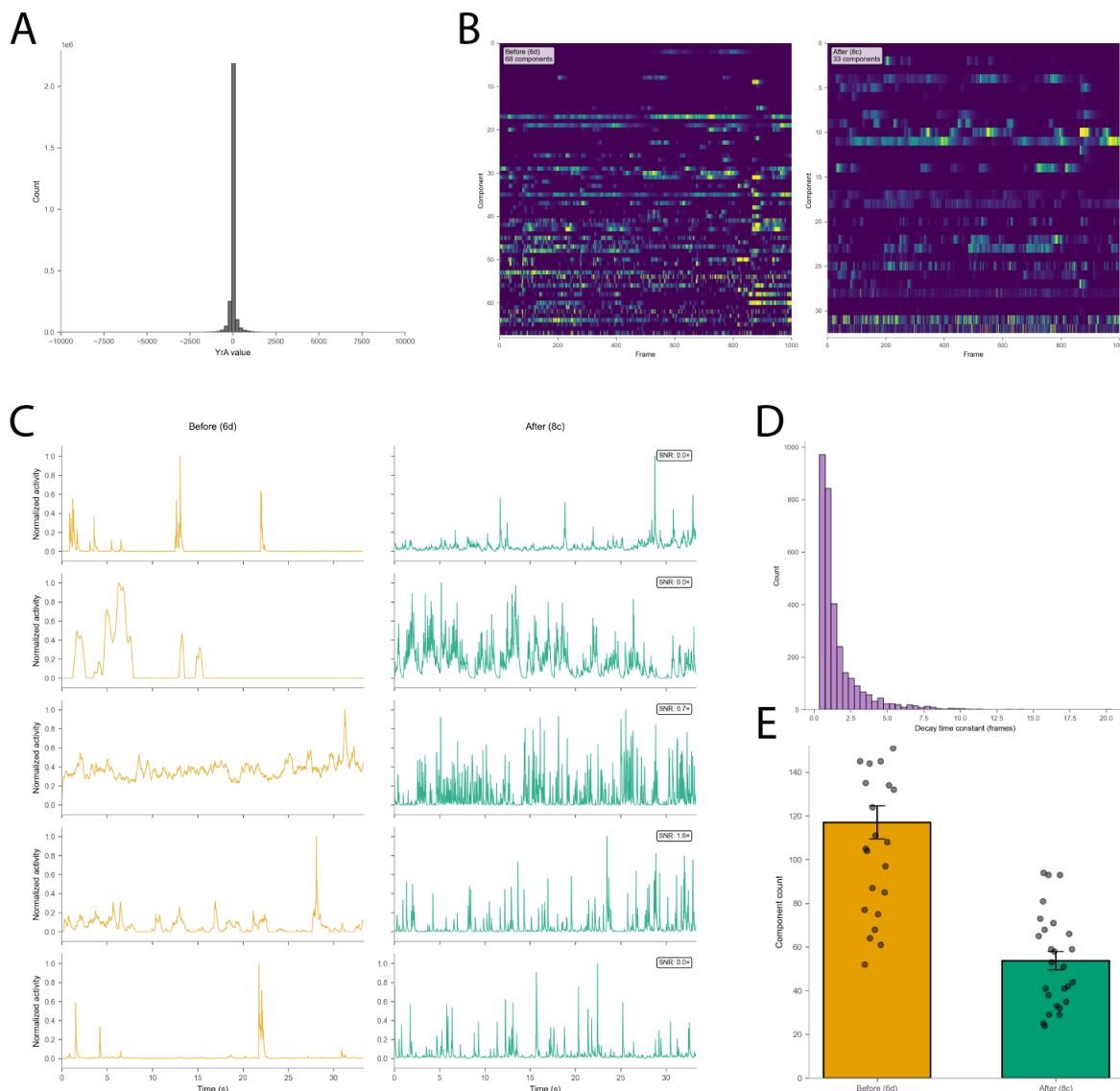


Figure 2.7: Final temporal refinement with updated spatial footprints produces high-quality curated neural components. **(A)** Distribution of YrA values (residual after projecting refined spatial components onto the movie) centered near zero, indicating successful demixing and background subtraction with improved spatial footprints compared to first temporal pass. **(B)** Heatmaps of temporal activity for all components for a single session before (left) and after (right) the final temporal update. Each row is one component. Color intensity represents normalized calcium activity across frames. **(C)** Representative example traces showing post-first-pass temporal update traces (left, 5 examples) and denoised calcium traces after the second temporal refinement (right, 5 examples). **(D)** Distribution of estimated decay time constants across all final components, showing median decay of 1.0 frames (mean = 1.6 frames, range 0.4-20.3 frames), consistent with first temporal pass. **(E)** Component retention across final temporal filtering, comparing counts before (left, yellow/gold bar, median = 117.1 ± 7.7) versus after (right, teal bar, median = 52.6 ± 4.2) quality control, representing 55.1% rejection rate.

2.2.8 Validation Against Synthetic Ground Truth (Figure 8)

We next asked whether MPS recovers known sources from data with a defined ground truth. We processed the synthetic recording at four durations of 10, 20, 60, and 180 minutes. Recovered footprints were matched to ground-truth cells by centroid distance, and recovery was scored by spatial area ratio and by the temporal correlation between each matched pair (Figure 2.8).

At every duration, MPS recovered all 100 ground-truth neurons with a precision and recall of 1.0 (F1 of 1.000). Spatial agreement was high across durations. The mean area ratio was 0.984 at 10 minutes, 0.900 at 20 minutes, 0.892 at 60 minutes, and 0.911 at 180 minutes (Figure 2.8D). Temporal fidelity was also high. The mean per-neuron r^2 between recovered and true traces was 0.987, 0.985, 0.985, and 0.982 for the same durations (Figure 2.8E). Recovery did not degrade as the recording lengthened, showing that MPS stays accurate over multi-hour sessions.

The 180 minute run is shown in detail. Recovered and true footprints overlapped closely across the field of view (Figure 2.8A). A single-neuron zoom of the weakest spatial match still showed strong agreement, with a temporal correlation of 0.993 and an area ratio of 0.808 (Figure 2.8B). Even the three matches with the lowest temporal correlation tracked the true traces closely (Figure 2.8C). Finally, we profiled resource use. Peak memory and peak runtime were recorded for each MPS step (Figure 2.8F–G).

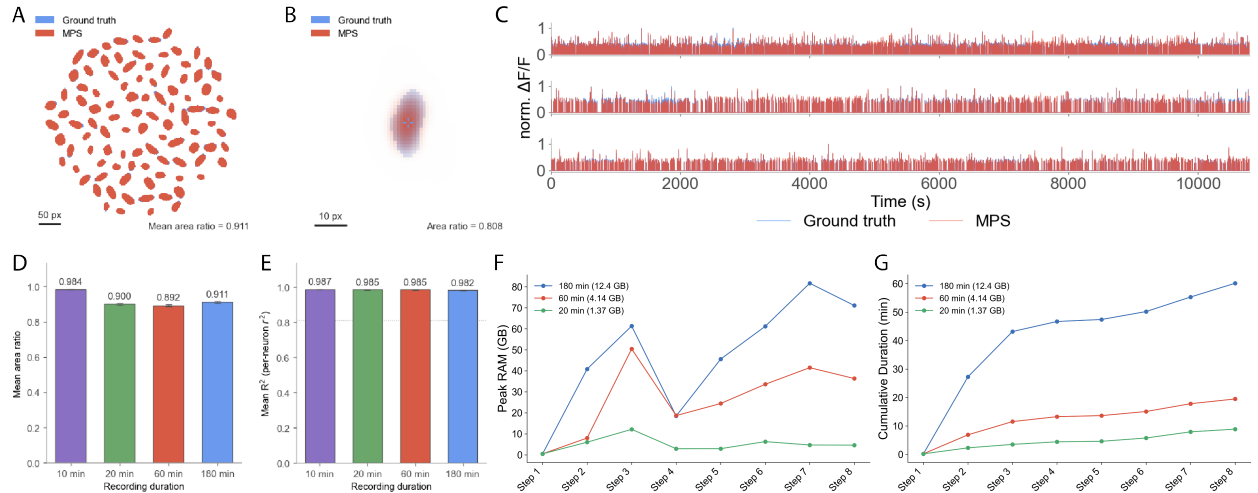


Figure 2.8: MPS recovers synthetic ground-truth sources with high spatial and temporal fidelity across recording durations. **(A)** Spatial overlap of ground-truth and MPS footprints for the 180 minute recording. Ground-truth footprints and MPS footprints are shown together across the cropped field of view. **(B)** Single-neuron zoom of the weakest spatial match. Markers show the ground-truth and MPS centroids. **(C)** Three representative matched traces with the lowest temporal correlation, normalized fluorescence over time, with ground truth and MPS overlaid. **(D)** Mean spatial area ratio across recording durations of 10, 20, 60, and 180 minutes. Error bars show SEM across matched neurons. **(E)** Mean per-neuron r^2 between recovered and true traces across the same durations. The dotted line marks an r^2 of 0.81, equivalent to $r = 0.9$. Error bars show SEM across matched neurons. **(F)** Peak memory use for each MPS step. **(G)** Peak runtime for each MPS step.

2.2.9 Summary

MPS combines (i) interactive cropping to avoid analyzing the uninformative parts of the video, (ii) NNDSVD to provide nonnegative, interpretable initial factors without unconstrained least-squares, and (iii) parallel, out-of-core temporal (AR(2), CVXPY) and spatial updates, producing accurate, scalable analyses of multi-hour recordings on standard lab hardware. The integrated Data Explorer closes the loop with transparent manual QC, supporting rigorous and reproducible workflows.

2.3 Discussion

2.3.1 Accessibility

A primary advantage of MPS is its accessibility to users of varying computational backgrounds. Unlike prior calcium imaging pipelines that require deep programming knowledge or command-line operation (e.g. CaImAn, MIN1PIPE, and CNMF-E in Python/MATLAB, or Minian via Jupyter notebooks), MPS provides a fully graphical user interface that guides the experimenter through each analysis stage. All major functions can be executed with menu selections and button clicks, with no need to modify code. All code is open-source, with detailed comments to facilitate transparency and ease of use. This GUI/no-code design dramatically lowers the barrier to entry. Users who are unfamiliar or intimidated with coding can now easily access miniscope processing technology through MPS, democratizing the analysis of multi-hour long miniscope data (Dong et al. (2022b)). The software includes built-in tooltips, tutorials, default parameter presets, full automation of processing, and a data explorer platform further simplifying usage for beginners. For users wishing to explore parameter behavior on manageable data, a representative sample dataset with annotated analysis code is publicly available at https://github.com/ariasarch/MPS_Sample_Code, providing a reproducible starting point without requiring access to the full multi-terabyte recordings used for benchmarking. Thus, what was once a scripting-intensive workflow was transformed into an interactive application accessible to the broader scientific community via MPS. This improved usability addresses a key limitation of earlier pipelines, wherein hard-coded parameters and lack of user guidance often made it difficult to trust and validate results without extensive tinkering. By transforming advanced calcium imaging analysis into a point-and-click process, MPS allows researchers to focus on experimental questions rather than computational setup and system management. Additionally, these results can be easily repeated or shared since the GUI enforces a consistent sequence of steps and settings, in contrast to ad-hoc custom scripts. Overall, the accessibility of MPS is poised to empower many labs to adopt miniscope imaging that previously found the data analysis too daunting (Freeman (2015); White et al. (2020); Aharoni et al. (2019)).

2.3.2 Scalability for Longer Recordings

Beyond usability, MPS achieves scalability rarely attainable in existing miniscope analysis software. Benchmarking across sessions up to three hours demonstrated stable performance and predictable

memory behavior even on constrained workstations. By combining out-of-core Dask execution, blockwise CNMF updates, and efficient chunking of intermediate arrays, MPS processes continuous videos exceeding 95,000 frames (~ 2.7 hours) without requiring manual chunking or external cluster resources. This is a significant improvement over prior implementations that saturate memory or fail during motion correction. Runtime scales linearly with dataset duration, and the memory footprint remains bounded by user-defined caps. Even when capped at 200 GB of memory, the full pipeline completed processing of 7.26 TB of miniscope data (total 77 hours of raw video) within 55.6 hours on a single Xeon workstation. This scalability ensures that experiments demanding high temporal coverage or long-term tracking can now be analyzed on ordinary lab hardware, promoting reproducibility and widespread adoption.

2.3.3 Extensibility to New Modalities

MPS was intentionally designed with modularity, allowing extensions beyond calcium imaging. Each analysis step is encapsulated as a standalone module within the pipeline framework, with standardized data interfaces that make it straightforward to integrate new frontends or replace existing algorithms. For example, future versions could incorporate voltage imaging, fiber photometry, or photometric motion tracking modules. Similarly, deep-learning models for ROI detection or denoising can be plugged into the pipeline through the model integration interface, which already supports loading pre-trained models from external frameworks such as PyTorch or TensorFlow. This modularity ensures that MPS can evolve alongside new imaging modalities and community-driven innovations while retaining the same no-code workflow.

2.3.4 Limitations

While MPS expands accessibility and scalability, several limitations remain. The current design focuses on preprocessing and CNMF-based source extraction; online or real-time decoding is not yet implemented. Additionally, although Dask parallelization enables large-scale computation on a single workstation, true distributed or cluster-scale processing remains future work. GPU acceleration, while feasible for some submodules, is not yet uniformly implemented due to hardware-specific dependencies. Future updates may include optional CUDA-accelerated paths and additional preprocessing algorithms to further improve performance. Another limitation is that long recordings with highly dynamic fluorescence or photobleaching may require customized baseline correction routines not yet part of the GUI presets. Nevertheless, these limitations represent tractable engineering improvements rather than conceptual barriers.

2.3.5 Community Impact

The primary impact of MPS lies in democratizing advanced calcium imaging analysis. By enabling non-programmers to perform analyses that were previously accessible only through scripting, MPS lowers the technical threshold for neuroscience research. This in turn fosters inclusivity and

reproducibility, aligning with broader open-science principles. The code base, released under an open-source license, encourages transparency and collaboration, while the GUI ensures consistent step ordering and parameterization across users. These design choices promote a unified standard for miniscope data analysis that can serve as a foundation for future benchmarking and cross-lab validation studies. Ultimately, MPS empowers the neuroscience community to shift focus from technical barriers to scientific discovery, accelerating progress in systems neuroscience.

2.4 Methods and Materials

2.4.1 Dataset Description and Scope of Validation

All validation and benchmarking results reported in this manuscript are based on a single experimental dataset consisting of 30 long-duration miniscope recording sessions collected from a single experimental paradigm. Recordings were obtained from 8 Long-Evans rats (4 male, 4 female) implanted with GRIN lenses (0.6 mm diameter, 6.3 mm length; GRINtech) targeting the right lateral habenula (AP -3.65 mm, ML +0.85 mm, DV -5.2 mm for males; scaled $\times 0.95$ for females). Animals expressed jGCaMP8f via AAV1-EF1 α -jGCaMP8f-WPRE. Sessions were acquired using a Miniscope V4 system at 10 Hz with a field of view of approximately 600×600 pixels. Each session spanned 2.4–2.9 hours (median 2.7 hours; 86,713–102,874 frames), yielding 30 total sessions, approximately 77 hours and 7.26 TB of raw video data. All sessions used the same calcium indicator, optical configuration, and acquisition parameters. All recordings were processed independently, with no information shared across sessions during any pipeline stage. Accordingly, the analyses presented here are intended to evaluate the performance, stability, and scalability of MPS in a controlled long-duration regime, rather than to claim universal parameter generalizability across brain regions, indicators, or optical systems. Users working with different brain regions, calcium indicators, optical configurations, or acquisition parameters should treat the parameter values reported here as starting points requiring iterative validation against their own data.

2.4.2 Synthetic Ground-Truth Validation Dataset

To evaluate source extraction against a known reference, we generated a fully synthetic calcium imaging dataset with specified neuron locations, footprints, and activity. The generator and the degradation model are openly available at https://github.com/ariasarch/MPS_Sample_Code. This dataset is a reproducibility and validation resource and is separate from the 30 session benchmark. All random draws use a fixed seed, so the dataset is exactly reproducible.

The synthetic recording matched the acquisition parameters of the real data. It contained 100 neurons in a 600×600 pixel field of view at 10 Hz over a 3 hour session (108,000 frames). Neurons were placed by rejection sampling within a central region of interest of radius 170 pixels. Somatic radii ranged from 4 to 8 pixels. A minimum center-to-center spacing of 25 pixels and a 10 pixel border margin were enforced. Spatial footprints were modeled as irregular ellipses. Each footprint used a log-normal axis ratio with a median near 2.15 and bounds of 1.3 and 4.0, a random orientation, and

a low-frequency angular wobble on its boundary. Each neuron’s fluorescence followed a second-order autoregressive GCaMP8f model with coefficients of 0.90 and -0.10 at 10 Hz. Spikes were drawn as sparse nonnegative events and passed through the autoregressive recursion under a nonnegativity constraint. Neurons were assigned to three activity tiers. High, sparse, and low tiers accounted for 30, 40, and 30 percent of cells and used progressively lower firing rates and amplitudes. Photometry used a baseline of 800 arbitrary digital units, a peak $\Delta F/F$ of 2.0, and per-pixel Gaussian noise with a standard deviation of 20 units. A short noiseless clip was also rendered as a clean reference.

The clean signal was then degraded to emulate a raw miniscope recording. Each frame added a diffuse neuropil haze formed from a blurred mean of neural activity and scaled by 0.15. A slow spatial vignette was added as a smoothed background with a peak-to-peak amplitude of 150 units. Independent Gaussian sensor noise was added at every pixel. Motion was introduced as a slow first-order autoregressive drift bounded at 3 pixels, combined with sparse transient jumps of up to 12 pixels that decayed over 8 frames. Frames were shifted by whole-pixel translation, and the wrapped borders were filled with the baseline value. The degraded video was written as 8-bit clips after mapping the 12-bit sensor range onto the 0 to 255 range. The ground-truth motion trace was saved alongside the clips. Because every footprint, trace, spike train, and motion vector is known exactly, this dataset supports direct measurement of recovery accuracy for spatial and temporal components. It complements the real 30 session benchmark, which characterizes scale and runtime but provides no ground truth. Users can regenerate the dataset or vary its parameters to probe pipeline behavior on tractable data before processing full recordings.

2.4.3 Ground-Truth Recovery Analysis

To quantify recovery, MPS output was compared against the known synthetic sources. The pipeline was run on the degraded video at four durations of 10, 20, 60, and 180 minutes, and the final spatial and temporal matrices were loaded for each run. Recovered footprints were first screened for duplicates. Two components were treated as duplicates when their spatial cosine similarity was at least 0.95 and their centroids fell within 5 pixels, and the redundant copy was set aside. Frames removed during preprocessing were restored in the temporal traces by linear interpolation over the retained frames, so that recovered and true traces shared a common time base. Recovered components were matched to ground-truth cells by centroid distance. A global translation between the two sets was first estimated as the median offset of an initial optimal assignment. Components were then paired by solving a linear assignment problem on the centroid distances, and a pair was accepted when its centroid distance fell within 15 pixels. Precision, recall, and F1 were computed from the accepted matches. For each matched pair, spatial agreement was summarized by the area ratio. Footprints were thresholded at 10 percent of their peak, and the area ratio was the smaller thresholded area divided by the larger. Intersection over union and the Dice coefficient were also recorded. Temporal agreement was summarized by the Pearson correlation between the true fluorescence trace and the recovered trace over the shared frames. The reported r^2 is the mean of the squared per-neuron correlation across matched cells. Peak memory and peak runtime were

recorded per step during processing.

2.4.4 Step 0: Download

MPS provides comprehensive installation support across multiple platforms, ensuring broad accessibility for the neuroscience research community. The software is distributed through a dedicated installation website (https://ariasarch.github.io/MPS_Installer/) that offers platform-specific installers optimized for different operating systems. For macOS users, MPS is available in both ZIP and DMG formats, supporting macOS 12+ on both Apple Silicon and Intel architectures. The application is signed and notarized so that macOS does not block it at launch. The installer writes into a user-owned directory rather than a protected system path, since attempting the latter produces an rsync permission error and leaves the install incomplete. Installation instructions on the documentation site now address this case directly, and installation was re-tested on clean macOS systems under the revised instructions. Windows users can access a dedicated EXE installer compatible with Windows 10/11 (64-bit systems), requiring approximately ~1GB of disk space during the initial setup process.

The installation process is fully automated and consists of four key steps: (1) downloading the MPS payload from the GitHub repository, (2) creating appropriate shortcuts on Windows or a self-contained application on macOS, (3) establishing the Python environment during the first launch, and (4) enabling instant subsequent launches. This streamlined approach minimizes technical barriers for researchers while maintaining robust functionality. System requirements are modest: ~32GB RAM minimum (~64GB recommended), stable internet connectivity for initial setup, and the aforementioned disk space allocation. The software architecture prioritizes accessibility through its intuitive graphical interface while maintaining the computational power necessary for complex calcium imaging analysis. Support infrastructure includes comprehensive documentation, an active issue tracking system for bug reports and feature requests, and community discussion forums (https://github.com/ariasarch/MPS_1.0.0/discussions/). The complete source code is available on GitHub, enabling customization for specialized laboratory requirements and fostering collaborative development within the research community (https://github.com/ariasarch/MPS_1.0.0/). To support reproducibility without requiring access to the full 7.26 TB dataset, a representative sample dataset and accompanying analysis code are publicly available at https://github.com/ariasarch/MPS_Sample_Code. A detailed parameter selection walkthrough, including guidance on interpreting diagnostic outputs at each pipeline stage, is also hosted there for users wishing to explore parameter choices on tractable data.

Table 2.2: Variables Produced by Each Step of the Miniscope Processing Suite (MPS).

Variable	Formats	Description	Step
processing_parameters	json	All pinned parameters for reproducibility	Global
processing_stats	json	Runtime, memory, and component counts per step	Global
Y	zarr	Raw video data used for preprocessing	2a
chunking_info	json	Chunk layout used by Dask for parallelization	2a
line_splitting_frames	txt	Frames flagged for line-split artifacts	2a
dropped_frames	txt	Frames removed after QC filtering	2d
motion_xy	zarr	Per-frame motion vectors ($\Delta x, \Delta y$)	2c
Y_hw_chk, Y_fm_chk	zarr	Validation stacks after motion correction	2e
Y_hw_cropped	zarr	Cropped field-of-view for ROI selection	3a
A_init, C_init	zarr	NNDSVD-initialized spatial/temporal matrices	3b
A, C (NNDSVD)	zarr	Decomposed factors after NNDSVD initialization	3b
b, f	zarr	Background spatial/temporal components	3b
nndsvd_results	dir	Diagnostics and spectra for initialization	3b
temporal_components	zarr	Candidate temporal signals for masks	4d
spatial_components	zarr	Candidate spatial masks before CNMF	4d
A_merged, C_merged	zarr	Merged components after initialization	4g
A_merged_coords	json	Unit coordinates and geometry metadata	4g
AC_backup	dir	Checkpoints of cleaned A/C before CNMF	4f
sn_spatial	numpy, json	Spatial noise estimates per pixel	5a
A_filtered, C_filtered	zarr, numpy	Validated components after QC	5b
C_filtered_coords	json	Metadata of surviving components	5b
YrA	zarr, numpy, json	$A^T Y$ projections for temporal updates	6a
temporal_params	json	Temporal update parameters (AR order, λ)	6a
A, C (new/filtered)	numpy	Updated temporal matrices after QC	6b
b0, c0, g	numpy	Baseline and autoregressive coefficients	6b
temporal_results_summary	json	Summary statistics of temporal refinement	6c
dilation/boundary summaries	json	Region growth and boundary statistics	7a–7c
clusters, cluster_bounds	json, pickle	KD-tree clustering and bounding boxes	7b–7c
component_valid_mask	numpy	Boolean mask of valid components	7b
multi_lasso_results	pickle	Multi-lasso regression diagnostics	7e
A_updated	zarr	Spatial footprints after regression updates	7e
A_merged (raw/smooth)	zarr	Reconciled spatial maps after smoothing	7f
merging_results_summary	json	Overlap merges and area metrics	7f
YrA_updated	zarr	Recomputed projection with refined A	8a
b0_final, c0_final, g_final, f_final, S_final	zarr	Final baseline, AR coeffs, background, spikes	8b
coords, processing_stats	json	Final geometry and run statistics	8c

2.4.5 Step 1: Graphical User Interface, Environment Provisioning, and Automation

MPS is implemented in Python (tested on Python 3.8.15) and distributed as a standalone, point-and-click application. Dependencies are explicitly pinned at installation time to ensure reproducibility. A conventional environment-based installation pathway (PyPI and/or Conda) is under development as an additional distribution option for users preferring standard package management workflows. Independently of packaged distribution, each pipeline step is also exposed as an importable Python module with a stable programmatic interface, allowing advanced users to script, batch, or headlessly run the complete workflow without the graphical interface. The main application imports all step definitions at launch and exposes them through a centralized controller, ensuring that the GUI can orchestrate modules consistently across runs. Additionally, all analysis parameters are centrally managed and saved to a single JSON configuration, enabling exact reproducibility and seamless resumption. The configuration captures both global settings (e.g., worker count, memory limits, and dataset paths) and per-step options (e.g., segmentation thresholds, NNDSVD rank, deconvolution hyperparameters). Representative keys are included and persisted alongside the dataset metadata for easy referencing.

Upon launch, the GUI presents a stepwise workflow with a status bar and progress indicator. Users can navigate the pipeline incrementally (Previous/Next) or operate it in fully automated mode. Moreover, a background process checks for current processing to protect the user from moving from step to step prematurely and risking a loss of data. Automation controls are exposed via the Automation menu allowing hands-off execution from any stage. An autorun indicator in the navigation bar reflects the current automation state, and the default delay between steps is user-settable. Users are advised to gain familiarity with the software and its parameters before enabling automation to ensure appropriate configuration. Parallel execution is handled automatically by the system; however, advanced users can access the Dask Dashboard via the ‘Open Dask Dashboard’ option, which provides a live link for monitoring task progress, memory usage, and worker activity. Moreover, for iterative work, a Load Previous Data dialog assists in restoring prior sessions detecting available cache artifacts (e.g., .zarr, .npy, .json), lets users select which to load, and can mark steps “completed through” a chosen stage so that processing resumes exactly where intended in case of an unexpected session termination. Parameters are auto-saved on exit to prevent configuration drift between runs as well.

Together, these design choices such as, environment provisioning via a one-click launcher, centralized JSON parameterization, menu-level automation toggles, optional notifications, and live dashboard access lower the operational complexity of running a full CNMF pipeline while preserving rigorous control over analysis settings.

2.4.6 Step 2: Data Preprocessing

Step 2 prepares raw miniscope video data through a series of cleaning operations to enhance signal quality and reduce artifacts prior to source extraction. This preprocessing pipeline is subdivided into Steps 2a–2f, each described below. For all subsequent cost analysis, let T be the number of frames in the recording and M the number of pixels per frame.

Step 2a: Video Loading and Format Standardization

The pipeline begins with loading miniscope videos via OpenCV (with FFMpeg used only to probe stream metadata) and materializing them as an xarray backed by Dask (lazy, chunked array), enabling out-of-core processing instead of holding full movies in RAM. The loader uses a regex file-pattern to select inputs and, in the current build, defaults to the AVI video format commonly saved by the Miniscope V4. For each file, MPS fprobes the stream and record width, height, and frame count and log the file path and size for traceability. Frames are read through a delayed OpenCV decode (cv2.VideoCapture), written to a chunked store, and wrapped into a Dask array of dtype uint8 with dimensions [frame, height, width]. The resulting array can be optionally cached on disk as a chunked store for reuse in later steps. An optional automated screen flags frames consistent with the “line-splitting” artifact. In the current implementation, the mean intensity is computed for every frame in the leftmost 20 columns and mark frames whose left-edge mean exceeds a robust threshold (global mean + $2 \times$ the Standard Deviation across frames):

$$\text{threshold} = \mu_{\text{global}} + 2\sigma_{\text{frames}} \quad (2.1)$$

Indices of flagged frames are saved, logged, and removed from the xarray before downstream processing. This straightforward yet effective safeguard is absent from existing pipelines, underscoring MPS’s emphasis on practical quality control. This built-in QC prevents obviously corrupted frames from propagating into motion correction and demixing. Computationally, byte I/O and decoding scale linearly with dataset size, and thus, the total cost is $O(M \times T)$. The line-split screen adds a small constant-factor pass over a fixed-width strip ($\sim 20 \times$ height per frame).

Line splitting is fundamentally a data-transport artifact rather than an optical or sensor-physics phenomenon. In the Miniscope V4, pixel data from the global-shutter PYTHON 480 sensor is read out row-by-row and streamed as a ~ 1 Gbps FPD-Link III serial signal across a micro-coaxial cable and a rotating commutator slip ring before deserialization and USB packetization (Aharoni and Hoogland (2019b)). A transient interruption anywhere along this chain — most commonly mechanical instability of the commutator slip ring, loss of serializer–deserializer phase-locked-loop lock, or a DMA buffer transition in the USB controller — corrupts precisely the rows in transit during that window, circularly shifting them or stitching them from an adjacent temporal frame. Because the standard DAQ firmware carries no CRC, parity, or retransmission on the pixel data path, these events reach disk undetected. The horizontal shift is deterministically related to the disruption duration Δt and the pixel clock f_{pix} ($\delta = \Delta t \cdot f_{\text{pix}}$, i.e. tens of pixels per microsecond at

the PYTHON 480’s readout rate), and its interaction with the objective’s radial vignetting produces the characteristic brightness step at the seam. Notably, this is not a rolling-shutter motion artifact: even the global-shutter sensors used in the V4 read out sequentially from in-pixel storage, so any link-level glitch maps onto specific rows regardless of shutter type.

MPS exposes two frame-level detectors for this signature. The default (*basic*) detector exploits the left-edge brightening produced by the vignetting mismatch: it computes the mean intensity of the leftmost 20 columns of each frame and flags frames whose left-edge mean exceeds the robust threshold defined above ($\mu_{\text{global}} + 2\sigma_{\text{frames}}$). An optional (*advanced*) detector instead scores each frame by the standard deviation of its Gaussian-detrended row-intensity profile within the fluorescent signal zone where a sharp discontinuity at the split boundary inflates this residual and separates artifact from clean frames using an automatic gap-based threshold on the score distribution.

Although strip-wise phase correlation can localize the boundary row and recover the sub-pixel block shift, and a vignetting-aware remapping can restore both spatial alignment and photometric consistency (Guizar-Sicairos et al. (2008)), MPS deliberately flags and *excludes* affected frames rather than reconstructing them. Principled correction presupposes an accurate per-block shift estimate and a well-estimated flat-field (vignetting) model, and it cannot recover rows lost to complete link failure (true non-image data). This represents an imperfect reconstruction risks injecting structured error into the very spatial footprints and temporal traces that downstream source extraction depends on. Thus, because line splits are rare in practice (median 0.37% of frames; Figure 2.2I), simple exclusion removes the corruption with negligible loss of temporal coverage over multi-hour recordings, whereas in-painting would trade a well-characterized, localized data loss for a poorly-characterized systematic bias. We therefore treat line-split detection as a quality-control gate, not a restoration step, and leave optional vignetting-aware correction as future work for datasets where frame loss is non-negligible.

Step 2b: Background Subtraction and Denoising

Miniscope recordings often suffer from uneven illumination and pixel noise. MPS addresses these by first subtracting a static background image from the video. Concretely, a per-pixel minimum projection over all frames is computed to estimate the persistent background fluorescence (vignetting and baseline autofluorescence) under the assumption that each pixel intermittently approaches baseline during the session (Giovannucci et al. (2019a)):

$$B_{\text{static}}(x, y) = \min_t [I(x, y, t)] \quad (2.2)$$

This background image is subtracted from every frame, removing constant offsets while preserving relative calcium fluctuations:

$$I_{\text{bg-sub}}(x, y, t) = I(x, y, t) - B_{\text{static}}(x, y) \quad (2.3)$$

Users can choose among median filtering (robust to salt-and-pepper noise), Gaussian (smooths gradual pixel-level variation), bilateral (preserves edges while denoising), or anisotropic diffusion (iterative edge-preserving smoothing for stronger noise regimes), giving flexibility to tailor denoising to different noise profiles and experimental conditions. The most common option recommended is a frame-wise median filter (default 7×7 kernel) which suppresses high-frequency salt-and-pepper noise without blurring soma edges (Giovannucci et al. (2019a)):

$$I_{\text{denoised}}(x, y, t) = \text{median}[I_{\text{bg-sub}}(x', y', t) : (x', y') \in \mathcal{N}_{7 \times 7}(x, y)] \quad (2.4)$$

Finally, a white top-hat (opening-based background removal) is performed with a disk-shaped structuring element with the radius matched to typical soma radius; e.g. 15 pixels to estimate and remove diffuse neuropil/background glow (Lu et al. (2018)):

$$F_{\text{neuropil}}(x, y, t) = (I_{\text{denoised}}(x, y, t) \ominus B_r) \oplus B_r \quad (2.5)$$

$$I_{\text{MBR}}(x, y, t) = \max(I_{\text{denoised}}(x, y, t) - F_{\text{neuropil}}(x, y, t), 0) \quad (2.6)$$

This “rolling-ball” equivalent operation has a long history in biomedical image processing and was introduced to miniscope pipelines in MIN1PIPE (Sternberg (1983); Lu et al. (2018)). Each filter operates independently per frame, so the total cost is $O(M \times T)$ and parallelizes across CPU workers. At the end of Step 2b, the data now has a near-zero baseline and reduced high-frequency noise, improving robustness of downstream motion correction and cell detection.

Step 2c: Motion Correction by Cross-Correlation

Head-mounted miniscopes often exhibit motion artifacts due to physical movement of the brain in the skull. In Step 2c, MPS performs rigid motion correction on the preprocessed video using template matching with normalized cross-correlation computed in the Fourier domain (Lewis (1995)). A reference image is defined (by default, the mean projection after background subtraction; alternatively, a max projection can be used for higher contrast). For each frame, the 2D normalized cross-correlation is evaluated with the reference to obtain the translation $(\Delta x, \Delta y)$ that maximizes similarity:

$$(\Delta x_t, \Delta y_t) = \arg \max_{x, y} \mathcal{F}^{-1} \left[\frac{\mathcal{F}[I_t] \cdot \mathcal{F}[\mathcal{T}]^*}{|\mathcal{F}[I_t] \cdot \mathcal{F}[\mathcal{T}]^*|} \right] \quad (2.7)$$

To handle large displacements efficiently, templates are built recursively: the recording is split into short temporal chunks (default 3), each frame is aligned to its chunk-level template, and templates are combined hierarchically into a global reference. Sub-pixel accuracy is achieved by fitting the correlation peak via an upsampled DFT around the maximum (Guizar-Sicairos et al. (2008)). Frames may optionally be flagged as unreliable (blur, occlusion) using a shape/circularity check on the correlation map, which is disabled by default. The FFT-based correlation for one frame costs $O(M \log M)$; over T frames the total is $O(T \times M \log M)$. Parallelization across frames (CPU

workers) is used to accelerate this step. The output is a per-frame shift vector applied in Step 2e to remap the movie into a stable coordinate frame. For context, non-rigid/piecewise-rigid alternatives (e.g., NoRMCorre) are common in calcium imaging (Pnevmatikakis and Giovannucci (2017)); here a rigid variant is adopted, which is sufficient after denoising and background suppression in Step 2b.

Step 2d: Detection of Erroneous Frames

After motion estimation, MPS automatically screens for frames with excessive inter-frame motion and optionally removes them. The detector consumes the per-frame x/y shift estimates produced during motion correction and flags frames in which shift magnitude is an outlier relative to the run’s typical motion:

$$\text{flagged if } |\Delta x_t| > \mu_x + \tau \sigma_x \quad \text{or} \quad |\Delta y_t| > \mu_y + \tau \sigma_y \quad (2.8)$$

Concretely, for each motion dimension present (vertical/height, horizontal/width), the mean and standard deviation of the estimated shifts is computed and mark any frame in which shift exceeds a user-tunable threshold factor (default $\tau = 5.0$) times the standard deviation from the mean. When the “Drop Erroneous Frames” option is enabled (default: on), the flagged frames are removed in tandem from both the motion dataset and the reference video array carried forward from Step 2b; the arrays are then rechunked to keep frame-chunk sizes well-formed for downstream Dask operations. MPS writes two audit files: a per-step list of dropped indices (`dropped_frames.txt`) and a combined list (`all_removed_frames.txt`) that merges Step 2a line-split removals with Step 2d motion outliers, adjusting indices back to the original frame numbering for reproducibility. This screening is linear in the number of frames i.e., $O(T)$ with a small constant factor. Because the underlying arrays are Dask-backed, the work is executed out-of-core and can leverage available workers allowing for multi-hour recordings be sanitized without ever loading the full movie into RAM, avoiding scheduler stalls from oversized task graphs, and preventing a handful of extreme-motion outliers from corrupting templates and biasing downstream CNMF.

Step 2e–2f: Frame Transformation and Remapping and Post-Validation of Preprocessing

Once the optimal shifts are determined, MPS applies a per-frame rigid warp to generate a stabilized movie. Each frame is translated by its $(\Delta x, \Delta y)$ estimate and resampled with sub-pixel interpolation. The GUI exposes these options directly and saves the corrected data as a Zarr dataset, with optional derivatives chunked for time-series or spatial access to support downstream steps. The operation is linear in data size ($O(M \times T)$) and parallelized across frames using Dask. Sub-pixel remapping follows standard image-registration practice (Thévenaz et al. (1998); Guizar-Sicairos et al. (2008)). After Step 2e, MPS provides a validation panel that runs lightweight checks on the motion-corrected video array before advancing to Step 3. Users can set a sample size for analysis (default 1000 frames; 0 means all frames) and enable three checks: (i) frame-intensity statistics, (ii) NaN/Inf detection, and (iii) an optional chunking comparison that benchmarks two storage layouts carried forward from

Step 2e (frame-chunked vs spatial-chunked). These checks operate on the already-produced Step 2e arrays and execute via Dask out-of-core computation, and thus, runtime scales approximately linearly with the number of frames evaluated ($O(T)$ for the chosen sample).

2.4.7 Step 3: Field-of-View Cropping and Dimensionality Reduction

To reduce computational load and focus analysis on relevant regions, MPS applies interactive cropping of the field of view followed by truncated NNDSVD for dimensionality reduction. This step removes non-informative pixels, accelerates downstream processing, and provides biologically interpretable initialization factors for CNMF

Step 3a: ROI Cropping

Miniscope videos often contain dark or irrelevant regions e.g., the circular vignette periphery or tissue outside the target area (Leong et al. (2003)). MPS provides an interactive cropping tool that lets the user define a focused field of view via a radius slider (default 75% of the smaller image dimension) and X/Y offsets, with “Preview Crop” and “Apply Crop” actions and a “Use Full Frame” reset. The preview overlays the proposed crop on a mean-intensity image; applying the crop saves chunked Zarr arrays for downstream steps and logs progress in the GUI. All pixels outside the selected region are discarded for subsequent analysis. This simple operation typically reduces pixel count by $\sim 70\%$ for a 0.6mm GRIN Lens (Inscopix-distributed GRINTECH optics), yielding a comparable speedup in later steps whose cost scales with the number of pixels M (e.g., a 70% reduction in M gives $\sim 3.3\times$ faster computation). Implementation is a per-frame mask-and-trim (rectangular crop aligned to the selected center), with total complexity $O(M \times T)$ and trivial parallelism across frames. After cropping, all frames are truncated to the crop’s bounding box, and M thereafter denotes the pixel count with this new ROI.

Step 3b–3c: Non Negative Double Singular Value Decomposition (NNDSVD) for Dimensionality Reduction.

To obtain a compact, nonnegative initialization for CNMF, the cropped, motion-corrected video dims is reshaped: [frame, height, width] to a 2-D matrix [frames, pixels] by stacking the spatial axes, then shift the data to nonnegativity by subtracting the global minimum and adding a small epsilon. An in-memory randomized SVD with user-selectable parameters (#components, default 100; power iterations, default 5). This yields U , S , and $V^\top$ where $U \in \mathbb{R}^{T \times K}$, $S \in \mathbb{R}^K$, and $V^\top \in \mathbb{R}^{K \times M}$ is then attempted:

$$Y \approx \sum_{k=1}^K u_k \sigma_k v_k^\top \quad (2.9)$$

From these, an NNDSVD is constructed by taking thresholded positive/negative parts of each singular triplet and (optionally) applying a spatial localization score to prefer concentrated spatial

maps:

$$A_{\text{init}}, C_{\text{init}} = \text{NNDSVD}(Y, K) \quad (2.10)$$

The first component is reserved as a background factor; the remaining components seed potential neurons. The resulting initial factors are written to disk as A_{init} [component, height, width], C_{init} [frame, component], and background components in space and time (b and f respectively) for the spatial and time background. (Boutsidis and Gallopoulos (2008); Halko et al. (2011)) Additionally, this step has a near-linear pass over the data with additional $O((M + T)K^2)$ terms, and thus, runtime scales with K components, frames, and pixels. After initialization, Step 3c provides a brief diagnostics and QA for the NNDSVD results. The GUI renders: (i) singular value spectrum and cumulative variance explained; (ii) a sparsity comparison (non-zeros in the raw SVD vs. the NNDSVD); (iii) an overlap map showing how many components cover each pixel; (iv) a background component panel (spatial map, temporal trace, histogram, and power spectrum); (v) per-component size distributions, relative singular values, max spatial intensities, and temporal variability; (vi) a component details view (spatial map, temporal trace, spatial profiles, and summary stats). Users can also export a CSV of component metrics for records/review.

2.4.8 Step 4: Initial Cell Detection and Component Initialization

Step 4 constitutes the initialization of potential cell components (spatial footprints and temporal traces) from the preprocessed video. MPS employs a two-pronged strategy for initialization: (1) image segmentation to directly detect cell locations and (2) matrix factorization-based initialization to decompose the video into the relevant components. By uniting direct image segmentation with matrix factorization, MPS introduces a previously unexplored initialization strategy that delivers an exceptionally robust and biologically faithful foundation for CNMF refinement.

Step 4a–4b: Peak Detection via Watershed Segmentation

To initialize potential cell footprints, MPS applies watershed-based segmentation to each NNDSVD spatial footprint from Step 3b. Each footprint is lightly smoothed with a Gaussian kernel (multi-scale search, $\sigma \in [0, 2]$ pixels) and local maxima are detected subject to a minimum inter-peak distance (default 20 pixels; searched over 10–30) and a relative intensity threshold (e.g., $> 10\text{--}20\%$ of the footprint peak). These peaks serve as markers for a marker-controlled watershed on the inverted projection, partitioning the image into candidate regions whose boundaries follow intensity saddles around each peak (Vincent and Soille (1991); Lu et al. (2018); van der Walt et al. (2014)). Parameters are biased toward slight over-segmentation at this stage (favoring sensitivity to dim or irregular somata) since redundant or irrelevant fragments are merged or dropped respectively downstream. The worst-case time complexity of the watershed transform is $O(M \log M)$ where M is the number of frames, but in practice is near-linear in the number of pixels for 2D images of this size. In the GUI, Step 4a performs a small grid search over `min_distance`, `threshold_rel`, and σ on a subsample to propose sensible defaults; results (including the best triplet and summary plots) are

written to state and auto-loaded by Step 4b, which then applies marker-based segmentation across components with optional minimum-region-size filtering and parallel execution. These controls are exposed directly in the Step 4a/4b panels (e.g., Run Parameter Search, Run Segmentation, Skip Background, Use Parallel Processing) and their choices are persisted to the session configuration for exact reproducibility.

Step 4c: Spatial Merging of Candidate Segments

MPS merges spatially proximate candidate segments from Step 4b using simple, size-aware centroid heuristics since the watershed segmentation usually produced a drastic overestimation of neural components. The step first loads the component list produced by Step 4b and optionally filters out very small segments (default min size = 9 pixels). The user controls four parameters in the GUI: Distance threshold (default 25 pixels), Size-ratio threshold (max larger/smaller, default 5.0), Minimum size (default 9 pixels), and Maximum merged size (default 5000 pixels). Merging proceeds greedily as follows: for each seed component not yet processed, MPS initializes a group and scans subsequent components where a candidate is added to the group if (1) the Euclidean distance between centroids is less than the set distance threshold:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} < d_{\text{threshold}} \quad (2.11)$$

(2) the size ratio is less than the set size-ratio threshold, and (3) adding the new component would not exceed the maximum merged size. After groups are formed, components in each group are merged by pixelwise summation of their spatial images and the merged centroid is recomputed via a center-of-mass calculation:

$$\vec{c}_{\text{merged}} = \frac{\sum_i A_i \vec{c}_i}{\sum_i A_i} \quad (2.12)$$

In the worst case, pairwise screening of N components is $O(N^2)$ with typically a few-hundred candidates; this pass is fast in practice.

Step 4d: Initial Temporal Signal Extraction

Before any CNMF refinement, a conservative estimate of each candidate unit’s activity is needed to (i) verify that the segment actually carries calcium signal and (ii) provide stable seeds for subsequent updates. Thus, for every candidate region, a footprint-weighted region-average trace is formed, at each frame, from the fluorescence of all pixels inside that region’s mask (reducing to a simple mean for binary masks):

$$C_{\text{initial}}(k, t) = \frac{1}{|R_k|} \sum_{(x,y) \in R_k} I(x, y, t) \quad (2.13)$$

Stacking all masks as columns gives a sparse “mask matrix” that, when multiplied by the movie, produces all traces in one pass; each trace is then divided by its region area to yield comparable units. Additionally, the cost scales with the number of regions $\times$ pixels $\times$ frames, but because

masks are sparse pixels are only iterated that belong to at least one region. The effective cost is proportional to (total masked pixels) $\times$ (number of frames), and thus, in a typical field (≈ 200 regions, ≈ 50 pixels each), that's $\sim 10,000$ pixel-sums per frame and is far smaller than processing the full image. Therefore, the data is streamed in time chunks (default 10k frames) to bound memory, with optional parallelism across regions/frames for seed initialization.

Step 4e: Spatial and Temporal Matrix Initialization

With step 4e, the segmented footprints and their region-average traces are turned into consistent spatial/temporal matrices that downstream CNMF can use, while enforcing basic quality checks and reproducibility. Users can set (i) spatial normalization mode (none, max, L1, or L2), (ii) minimum component size (default 10 pixels), (iii) maximum number of components (0 = no limit), (iv) whether to skip any background component (ID 0), and (v) NaN/Inf checks. The surviving components are defined as the units and stack their 2-D footprints into a spatial tensor with shape [units, height, width] and their 1-D traces into a temporal matrix with shape [frames, units]. Coordinates (height/width) are inherited from the Step 3b template to keep all downstream arrays aligned. If spatial normalization is enabled, each unit's footprint is rescaled by the chosen norm (max, L1, or L2):

$$A_k^{\text{norm}} = \frac{A_k}{\|A_k\|_p}, \quad C_k^{\text{norm}} = C_k \cdot \|A_k\|_p \quad (2.14)$$

and its corresponding temporal trace is inversely rescaled so that the implied reconstruction magnitude (spatial $\times$ temporal) is preserved. Lastly, the cost is linear in terms of number of units and their pixel extents (stacking and per-unit normalization), plus lightweight I/O and logging. With a few hundred units, this step is quite fast in practice.

Step 4g: Temporal Correlation Merging

For step 4g, MPS performs a single-pass, correlation-guided merge of candidate components using the cleaned spatial/temporal matrices from Step 4f. The GUI exposes four controls: (i) temporal correlation threshold (default 0.75), (ii) spatial overlap threshold (fractional, default 0.30), (iii) maximum merged size in pixels (default 5000), and (iv) an optional cap on the number of components to consider (0 = no cap). The temporal matrix is materialized to a dense array ($K \times T$) and computes a full Pearson correlation matrix with `np.corrcoef`:

$$\rho_{ij} = \text{corr}(C_i, C_j) \quad (2.15)$$

The overlap is computed using the overlap (Szymkiewicz–Simpson) coefficient (Simpson (1960)): the count of intersecting pixels between two masks divided by the size of the smaller mask:

$$\text{overlap}(i, j) = \frac{|A_i \cap A_j|}{\min(|A_i|, |A_j|)} \quad (2.16)$$

This metric equals 1 when the smaller mask is entirely contained in the larger and is generally $\geq$ IoU (Intersection over Union). Unlike IoU, which penalizes size asymmetry, the overlap coefficient is designed to capture inclusion and containment, making it especially well-suited for detecting redundant or duplicate cellular masks where one footprint may be fully nested within another. For surviving pairs, the algorithm grows a merge group by unioning any other components whose temporal correlation with either seed exceeds the threshold, provided (a) each proposed member overlaps spatially (by the same rule) with all current members and (b) the estimated union size of the group’s masks does not exceed the maximum-size limit. Components accepted into a group are marked processed to avoid reuse. All remaining components become singleton groups. Building the correlation matrix is $O(K^2 \times T)$ in the worst case (after a full materialization of C); group formation uses set/overlap checks that are fast for “few-hundred” components. Default thresholds (temporal $r = 0.75$, spatial = 0.30) are chosen to be conservative and users can tighten/loosen them in the panel. However, these steps are for increasing convergence of the CNMF by providing a better starting point, thus it is better to include more components, than less in this situation.

2.4.9 Step 5: Pre-CNMF Noise Estimation and Validation

Prior to CNMF refinement, MPS generates a pixelwise noise map from residuals of the background model and validates inputs for integrity and size-based plausibility. These checks establish component-specific noise levels used in the CNMF updates and ensure robust inputs for subsequent updates.

Step 5a: Noise Estimation

To obtain a spatially resolved estimate of measurement noise prior to CNMF refinement, a per-pixel noise map is computed from the residuals of the background model. The input movie is first cropped and motion-corrected (Step 3a). The background is represented by a single spatial component (from Step 3b) and a corresponding temporal trace. A background movie is formed as their outer product, subtract it from the data, and obtain residuals that capture what the background model cannot explain:

$$R(x, y, t) = Y(x, y, t) - b(x, y) \cdot f(t) \quad (2.17)$$

The per-pixel noise level is then estimated as the temporal standard deviation of the residuals, yielding a two-dimensional “noise standard deviation” image:

$$\sigma_n(x, y) = \text{std}_t[R(x, y, t)] \quad (2.18)$$

This map highlights which regions of the field of view fluctuate strongly after background subtraction, providing a more realistic measure of signal quality than a single global estimate. To account for heterogeneity in actively fluorescing regions, users are allowed to scale the noise inside “active background” areas by a multiplicative factor (default 1.5). Active regions are defined by thresholding the background component using either the global mean, global median, or a custom threshold. Inside the resulting mask, noise values are inflated accordingly. This adjustment helps prevent

bright, slowly varying regions from being mistaken as low-noise simply because they dominate the background. For additional stability, the noise map can be Gaussian-smoothed, which reduces pixel-level spikiness while preserving broad spatial trends. Because the entire procedure is linear in both pixel count and frame count, and runs out-of-core with xarray/Dask, it scales efficiently to multi-hour recordings without exhausting memory.

Step 5b: Validation and Setup

Before the alternating CNMF updates, an integrity check of inputs and (optionally) size-based component filtering is performed. Optional checks scan A , C , and the noise map for NaN values and report global ranges (the full movie Y is skipped for memory reasons); an estimated memory footprint (sum of A , C , and noise-map nbytes) is logged to help users gauge resource demands. For footprint sanity checks, each unit’s area is computed as the count of strictly positive pixels in its spatial map; from these sizes summary statistics are reported and the counts falling below a user-selected minimum or above a maximum. If “Apply size filtering” is enabled, units outside the specified range are removed by indexing on the `unit_id` dimension.

2.4.10 Step 6: Temporal Activity Update (CNMF Temporal Refinement)

Similar to CNMF implementations (Lu et al. (2018), Giovannucci et al. (2019a), Dong et al. (2022b)), temporal activity is refined (C , plus sparse events S and background terms) while holding the current spatial maps (A) fixed. Importantly, proceeding with this temporal update first, unlike other CNMF implementations, drastically reduces total runtime since a cleaner C narrows the spatial search in Step 7. Additionally, all previous steps (initialization, seeding, merges, and noise estimation) exist as a warm start for the next CNMF updates to speed CNMF convergence and stabilize this temporal and next spatial pass.

Step 6a–6b: Computation and Validation of Projected Fluorescence

Given the current spatial footprints and background model, Step 6a computes the residual activity tensor by fitting the movie as approximately the sum of $A \cdot C$ and B , then carrying forward the projections needed for deconvolution:

$$\text{YrA} = A^\top (Y - B) \quad (2.19)$$

This leads to a marked increase in processing time by performing this calculation once instead of the current standard that performs the YrA computation for every component. In practice, the movie is reshaped into a two-dimensional array (frames $\times$ pixels), the background is optionally subtracted, values are cast to 32-bit floats to bound memory use, and any NaNs are replaced with zeros. The result is a dense projection of the movie onto the component footprints, which is then saved in chunked Zarr format for efficient downstream access. Additionally, a sanity-check was

added to the YrA outputs before deconvolution by computing correlations and descriptive statistics and visualizing example traces and residual panels.

Step 6c: Automated temporal parameter suggestion

While many existing tools require users to determine suitable parameters through trial and error, MPS provides automated suggestions to streamline setup, such as proposing conservative defaults for the AR model and sparsity based on Step 6a YrA, the Step 5a noise map, and a user-specified subset of components/frames. For reference, the AR order p is proposed from autocovariance structure (the coefficients g are estimated later during the solve):

$$g = \text{AR-coefficients from autocovariance}(C_k) \quad (2.20)$$

set a sparsity weight λ proportional to local noise, and (optionally) normalize traces. Selection can be random, best/worst SNR, or by size, and all suggestions are logged for reproducibility. These parameters are meant to be sensible starting points, not exact ground truth and the user is encouraged to tune them to their dataset.

Step 6d: Update temporal components

In this step, each potential cell is assigned a calcium trace and corresponding spike train by solving a noise-aware sparse deconvolution problem that respects autoregressive calcium dynamics. For every unit, the method balances two terms: fidelity to the observed residual trace and sparsity of the inferred spikes:

$$\min_{C_k, S_k} \|YrA_k - C_k\|_2^2 + \lambda \|S_k\|_1 \quad \text{subject to } C_k(t) = \sum_{i=1}^p g_i C_k(t-i) + S_k(t), C_k \geq 0, S_k \geq 0 \quad (2.21)$$

Non-negativity is enforced so that all outputs remain physically interpretable. The implementation uses CVXPY, an open-source Python library for formulating and solving convex optimization problems, with a formulation that explicitly introduces spike variables and ties them to calcium dynamics through an autoregressive operator, guaranteeing a convex and stable optimization (Diamond and Boyd (2016)). Several design features make this implementation robust and scalable. Each trace is normalized, noise is estimated locally from the spatial footprint, and autoregressive coefficients are estimated adaptively in a way that ensures stability even on long or noisy recordings. Drift and offset terms are modeled explicitly, so slow baseline changes do not contaminate the spike train. If one solver backend fails, the system automatically falls back to alternatives, greatly improving reliability across datasets.

The data is broken into overlapping temporal chunks that are distributed across many workers. Within each chunk, components are solved sequentially, and then chunks are reconciled to reassemble continuous outputs. This chunk-level parallelization ensures that many temporal windows are processed simultaneously, and thus, since the cost per trace is proportional to the number of frames,

the end-to-end runtime scales linearly with the size of the dataset while achieving high throughput in practice. Additionally, by performing the temporal update before refining the spatial footprints makes later spatial updates converge faster, reduces instability, and cuts down on redundant iterations.

Finally, the pipeline includes comprehensive convergence tracking. Metrics such as objective value, primal and dual residuals, solve time, and autoregressive coefficients are logged for each component. These can be saved, visualized, and inspected to verify that optimization is proceeding correctly. This level of transparency helps ensure that results are both reproducible and interpretable. Refer to step 8a for more information.

Step 6e: Temporal quality control and filtering

Summary metrics are computed and remove obvious artifacts. Defaults include thresholds on total spike activity (`min_spike_sum`), variance of c_i (`min_c_var`), and a minimal spatial-sum check to avoid degenerate A - C pairings. The UI provides quick inspection of individual components and saves both the full and filtered outputs with chunked, NumPy-backed export for fast reload in later steps. This temporal-first ordering materially decreases wall time for spatial refinement. With cleaner C , the localized regressions in Step 7 operate on smaller, better-conditioned tiles, require fewer passes, and parallelize across clusters without contention.

2.4.11 Step 7: Spatial Footprint Update (CNMF Spatial Refinement)

Following the temporal update (Step 6), MPS updates spatial footprints while holding temporal activity fixed. Thus, after performing the temporal update first, cleaner temporal traces reduce ambiguity in how pixels and time are coupled - ensuring a faster spatial solve convergence, requiring fewer passes, and allowing for parallel execution across independent neighborhoods. The spatial update is designed for high parallel efficiency, so the workload is distributed across workers without heavy synchronization.

Step 7a–7d: Dilation-Bounded Search, Neighbor Discovery, and Localized Regression

In Step 7a, each component’s support set is expanded conservatively by dilation with a small user-defined radius. This defines a growth mask Ω that bounds how far a footprint can spread, shrinking the search space and preventing spurious expansion:

$$\Omega_k = \text{dilate}(A_k, r_{\text{dilation}}) \quad (2.22)$$

Step 7b builds a KD-tree (a k-dimensional, space-partitioning data structure for fast nearest-neighbor queries) over footprint centroids to identify spatial neighbors. Nearby or overlapping units are grouped into clusters, and for each cluster a bounding box Λ is computed. These clusters define disjoint tiles, allowing extremely parallel updates because clusters with no overlap can be processed independently.

Step 7c enforces conservative size bounds with a new per-component maximum growth parameter: each footprint can grow at most γ times its original support:

$$|A_k^{\text{new}}| \leq \gamma \cdot |A_k^{\text{old}}| \quad (2.23)$$

If an updated mask exceeds this cap, only the strongest pixels by coefficient weight are retained. This prevents runaway growth and guarantees that tile updates remain stable and independent.

Step 7d manages the regression parameters. Within each tile, footprints are regressed against the temporal traces of all components in the local cluster (plus an optional background term). A nonnegative sparse regression then updates footprints inside the dilation mask using an L1 penalty to promote compact footprints:

$$A_k^* = \arg \min_{A_k \geq 0, A_k \in \Omega_k} \frac{1}{2T} \|Y - A_k C_k\|_2^2 + \lambda \|A_k\|_1 \quad (2.24)$$

For single isolated units this reduces to a simple per-pixel non-negative least squares (NNLS) formula where CVXPY is the tool and NNLS becomes the problem structure.

Step 7e–7f: Parallel Execution and Post-Update Merging

The spatial update is not a single “apply regression everywhere” operation; instead, it is a staged process designed to refine each footprint inside a bounded region, using sparse computation and explicit growth controls to keep results biologically meaningful and computationally efficient.

The function implements a multi-penalty sparse regression for every pixel inside a component’s dilation mask. Instead of fitting the entire field of view, the algorithm restricts attention to pixels flagged by the dilation mask Ω . This means that only a 5 to 20% fraction of the pixels in a tile are ever considered leading to an immediate several-fold speedup relative to dense updates. For each pixel in Ω , the local trace is extracted from the movie, normalized, and validated. Traces with extremely low variance or non-finite values are skipped to prevent numerical instability. Regressors are built from the temporal traces of the components in the cluster (and background signals). The solver then runs a multi-LASSO strategy: instead of choosing a single penalty, it evaluates a logarithmically spaced range of penalties, solves each with non-negativity enforced (a custom non-negative coordinate-descent solver, with LassoLars available as a fallback), and collects statistics about sparsity, residuals, and support. These multiple solutions are then combined into a single coefficient vector, weighted by residual quality. This approach is robust to over or under penalization and provides a distribution of solutions. The pixel coefficient is accepted only if it clears an activity threshold. Coefficients below threshold are discarded, which ensures that noise fluctuations do not accumulate into spurious spatial spread. For accepted pixels, the algorithm logs residuals, sparsity levels, and support frequency across penalties, building a per-pixel statistical picture of confidence. These diagnostics are aggregated into maps (support frequency, residual map, coefficient distribution) that can be inspected for quality assurance.

A novel safeguard is the maximum growth factor. After solving all pixels in a component’s

dilation mask, the number of active pixels is compared to the original footprint size. If the footprint has grown more than a user-defined multiple (default $\gamma = 3\times$), the component is trimmed back. Specifically, pixels are sorted by coefficient magnitude, and only the top $\gamma \cdot |S_{\text{old}}|$ pixels are kept guaranteeing that components cannot explode in size due to overfitting or correlated noise. The growth cap is enforced consistently in two places: during per-component refinement and again when results are written back globally. The sparse design means that every step operates on a small fraction of the image, making the effective runtime proportional to the number of masked pixels rather than the entire tile. On typical datasets, this achieves an order-of-magnitude speedup relative to dense regression, while also yielding cleaner, more compact footprints. Together, these design choices transform the spatial update from a global, ill-posed problem into a sequence of small, well-posed, and bounded tasks. This is what makes it possible for MPS to refine thousands of footprints in multi-hour movies on a single workstation without exploding runtime or memory.

Step 7f finally handles post-update reconciliation. Overlaps are recomputed, and units with high similarity are merged. Statistics on area, compactness, and growth-cap enforcement are logged, and results are written back with full provenance. Because clusters were already defined to minimize contention, this merging stage is lightweight and can be parallelized as well.

Overall, ROI cropping in Step 3, watershed seeding in Step 4, noise-aware validation in Step 5, and temporal-first refinement in Step 6 all serve to make Step 7 a local, bounded, and highly parallel task. The dilation masks, cluster tiles, noise-weighted regression, and growth-cap innovation together transform one large ill-conditioned regression into many small, well-posed problems that can be solved side by side. This design reduces analysis time and memory pressure, improves demixing, and most importantly guarantees reproducibility across runs.

2.4.12 Step 8: Final Temporal Refinement and Output

After spatial updates, MPS performs a final temporal pass using refined footprints to improve spike inference and calcium trace accuracy. Results are then exported for curation in the integrated Data Explorer, providing a consistent and reproducible endpoint for analysis.

Step 8a–c: Re-projection, Temporal Updates, and Final Filtering

After spatial refinement (Step 7), the projection of the movie is recomputed onto the latest footprints to obtain an updated YrA (the per-component fluorescence obtained by multiplying $A^\top$ with the movie Y) exactly the same as Step 6a. In Step 8b, with YrA recomputed from the refined footprints, a last pass of temporal inference is performed. As in Step 6, each component’s trace $c(t)$ is modeled with an auto-regressive (p) calcium dynamics prior and an L1 penalty on the innovation (spike) sequence, subject to nonnegativity. When a Dask client is available, Step 8b dispatches independent temporal chunks to workers to reconcile overlapped windows to avoid edge artifacts. The last sub-step assembles results for analysis and sharing. The UI loads any saved parameters (e.g., minimum footprint size, minimum SNR, minimum correlation, whether to include metadata, and the compression level), checks which spatial and temporal arrays are present, and detects whether

spike estimates and background components are available from prior steps. On completion of the export, MPS celebrates with a small “fireworks” animation implemented in Tkinter. When the animation window is closed, MPS immediately offers a modal “Processing Complete” popup guiding the user to open the Data Explorer or continue elsewhere - so the celebration doubles as a clear hand-off into curation.

The final temporal pass repeats the same objective as the earlier update but with cleaner A , so both deconvolution and spike inference are better conditioned. In other words: $c(t)$ is estimated that (i) explains YrA, (ii) is nonnegative, (iii) follows an $\text{AR}(p)$ calcium evolution, and (iv) is sparse in its increments. Because footprints are now demixed and localized, YrA is less contaminated by neighbors, so the AR fit stabilizes and fewer iterations are needed. Moreover, chunked time windows with slight overlap allow linear-time passes over T while avoiding boundary artifacts when merging. The Step 8 implementation uses the same utilities (`process_temporal_parallel`, `merge_temporal_chunks`, defaults via `get_default_parameters`) as Step 6 to keep behavior consistent across passes.

2.4.13 Interactive Data Explorer and Manual Curation

The Data Explorer is a post-processing application built explicitly for manual quality control by users. Its purpose is to let a trained human verify that each extracted footprint corresponds to an anatomically plausible ROI and that its calcium trace exhibits visible transients in the raw field - before results are accepted for downstream analyses. It is not an automated adjudicator; rather, it’s the standardized place where an expert inspects, accepts, or rejects components and documents those decisions.

The Explorer launches as a dark-themed app with a split layout: left is an interactive A view; right is a stack of per-neuron C traces. One can load A .npy and C .npy independently; the tool auto-detects and converts input shapes so that A is handled internally as (H, W, U) and C as (U, T) . Clicking on a neuron in the A view both highlights its footprint and moves its trace panel to the top of the right stack, surfacing key units for rapid inspection. During playback, the A view is modulated frame-by-frame by the current activity level $C[t]$, effectively “blinking” active cells so an expert can visually link transients in the trace to the underlying image. Default playback is 10 fps, adjustable from the menu. The right-hand trace pane shows each selected neuron in its own axis, with a movable time cursor and play/pause controls. Window length (seconds), playback step (seconds per tick), number of traces shown, and FPS are all configurable so an expert can tune the inspection workflow to the dataset. In the A view, the app can overlay neuron IDs, display a static max projection for orientation, and, when a neuron is clicked, report its estimated pixel area at a chosen fraction of the footprint’s peak intensity. These features are designed to answer the specific questions a human curator asks: “Does this footprint have a plausible size and shape?”, “Do the largest events in the trace manifest as visible fluorescence in the movie?”, and “Is this unit redundant with a neighbor?”

Curation actions are explicit and reversible. Cells can be deleted via the menu (applied on save), and an “Unhide Last” action restores the most recently removed unit in place. The app maintains

an “unsaved changes” state, and saving prompts the user to choose a destination and filenames; A and C are written as edited NumPy arrays (float32), leaving the original outputs untouched. Session metadata (e.g., animal & session labels, cache/export paths) is displayed at the top so curated files are clearly tied to the analysis run. These guardrails ensure that expert decisions are captured deliberately and reproducibly. To keep inspection fluid on long recordings, the trace view thins the x-axis to ~ 1 -Hz visual density by default, and playback uses an exponential intensity mapping for activity-weighted A rendering (an optional exponential trace smoothing is also available, off by default). The A composite is normalized for display and can be blurred slightly for legibility without altering the underlying data. Together, these choices favor fast human judgment without sacrificing fidelity.

After Step 8 produces final A and C , a user opens them in the Explorer, plays through the movie while watching selected traces, and prunes dubious units (vascular structures, edge artifacts, low-SNR remnants, duplicates). The curated A/C pair is then saved and carried forward to registration or behavior alignment. By centralizing this last mile of manual QC into a purpose-built tool, rather than ad-hoc scripts, the pipeline ensures that expert judgment is consistently applied and documented across datasets.

2.4.14 Parallel Computing Implementation and Scalability Considerations

A cornerstone of MPS is the ability to handle extremely large movies with millions of pixels per frame (M) and 100,000 frames or more (T) on a single workstation without requiring a cluster. Every processing-expensive step is written to run on chunked arrays with out-of-core execution through Dask and vectorized NumPy/BLAS kernels, so memory use is bounded by the chunk size rather than the full dataset.

Time-chunked execution ensures that preprocessing and CNMF (Steps 2, 6, and 8) operate in windows of 1,000–5,000 frames at a time. Each window is read, processed, and cleared before the next begins. As a result, increasing T increases wall-clock time rather than peak RAM. On-demand spatial tiling is used only when necessary. If the field of view or number of units makes a single spatial pass impractical, MPS splits the update into non-overlapping tiles by dilating footprints and clustering nearby units. Only footprints that cross borders receive small overlaps, avoiding the reconciliation overhead of always-overlapping patches.

The scheduler relies on Dask’s distributed engine with a small pool of local workers (default 8). Large intermediates such as partial $A^\top Y$ blocks are memory-mapped or written to disk, and tasks are short-lived to prevent buffer buildup avoiding the “giant graph” stalls that often occur in long notebooks. For example: 1) Motion correction uses FFT-based correlation in windows. 2) Temporal updates (Steps 6 and 8) run independently across units and windows. 3) Spatial updates (Step 7) are restricted to dilation masks within each tile and solved with noise-weighted non-negative regressions. Since the number of components K is much smaller than $M \times T$, runtime is still dominated by the data size.

Theoretical complexity can be summarized as $T \times M \log(M)$ for motion correction, $M \times T$ for

per-frame preprocessing, $\bar{a} \times K \times T$ for trace extraction, $K^2 \times T$ for correlations and merges, and $(M + T) \times K^2$ for NNDSVD. In a pessimistic bound with full frames, the motion term dominates and predicts $\sim 9\text{--}10\text{h}$ for a $\sim 3\text{h}$ video. In practice, Step 3 cropping reduces the effective pixel count M by 55–70%. This alone yields a $2.2\text{--}3.3\times$ speedup across nearly every step. Combined with moderate K values (hundreds rather than the upper bound of 800) and the near-linear scaling of parallel workers, the actual wall time falls in line with about $\sim 1.3\text{--}2.6\text{h}$ for $\sim 95,000$ -frame runs. As an example, consider a 500×500 field of view, about 2.5×10^5 pixels. With 100,000 frames and 800 components, the pessimistic calculation gives roughly 4.5×10^{11} operations from motion correction and another 2.2×10^{11} from NNDSVD, suggesting $\sim 9\text{h}$. After a 60% crop the pixel count drops to 1.0×10^5 , which cuts those terms by a factor of ~ 2.5 , immediately bringing runtime to $\sim 3.5\text{h}$. With parallel scaling across 8 workers and slightly smaller K , the observed $\sim 2\text{h}$ total for many datasets matches expectation.

Even though the workstation used for benchmarking had $\sim 512\text{GB}$ RAM, a hard cap of $\sim 200\text{GB}$ was enforced to emulate $\sim 256\text{GB}$ -class machines and keep usage predictable. Competing pipelines either required manual chunking and stitching or failed on multi-hour runs due to graph bloat or memory exhaustion, whereas MPS processed entire sessions in a single pass with bounded memory and continuous progress reporting.

2.4.15 Statistical Analyses

Metrics were computed per recording and summarized across recordings with robust statistics. Baseline reduction (%) was calculated as one hundred times the difference between the pre- and post-values divided by the pre-value, where the measure is the median fluorescence over all pixels and frames. Cell-to-surround contrast was quantified as the fold-change in robust dispersion comparing post to pre. Rigid motion was estimated using frame-to-template cross-correlation; residual shift was defined as the median magnitude of per-frame displacement with interquartile range (IQR). Long-term drift, expressed in pixels per hour, was the slope from an ordinary least-squares fit of displacement versus time. Erroneous frames were flagged by heuristic detectors for line-split patterning, excessive blur, and saturation spikes. The effect of quality control was measured as the change in frame–template correlation, expressed in standard-deviation (SD) units relative to the pre-QC distribution.

For cropping and dimensionality reduction, metrics were computed on a per-session basis and summarized across sessions using robust statistics, including medians and interquartile ranges. Pixel load reduction was defined as the percentage decrease in active pixels per frame before and after cropping, while changes in runtime and memory usage were expressed as multiplicative speedup factors and percent reductions relative to the uncropped baseline. Variance explained was calculated as the cumulative fraction of total variance captured by truncated NNDSVD at a given rank K , and the elbow point was determined as the smallest K at which the marginal gain in variance explained dropped below one standard deviation of the pre-elbow slope. Iteration counts for CNMF initialization were compared between NNDSVD-seeded runs and both random and unconstrained

least-squares baselines, with differences expressed as percent reduction in iteration number.

For cell detection, initialization, and pre-CNMF validation, component counts and areas were measured at each stage of the pipeline, beginning with watershed segmentation and progressing through mask merging, duplicate removal, and validation. Oversegmentation ratio was defined as the number of watershed masks divided by the number of curated components. Duplicate merging was quantified as the proportion of candidate pairs with both high temporal correlation ($\rho > 0.75$) and spatial overlap (Szymkiewicz–Simpson containment coefficient > 0.30) that were collapsed into a single component. Noise estimation was performed by computing pixelwise residual variance after subtracting a low-rank background, with optional Gaussian smoothing to stabilize variance estimates; for each component, noise level was summarized as the mean residual standard deviation across its pixels. Validation included checks for NaN and Inf values, size thresholds based on user-defined minimum and maximum component pixel counts, and removal of invalid traces. The impact of quality control was quantified as the change in frame–template correlation after excluding erroneous frames, standardized in units of the pre-QC distribution. All analyses were conducted within individual sessions before aggregation across sessions to avoid distortion from outliers.

All statistics were calculated on a per-session basis and then summarized across sessions, with the session as the unit of analysis. Component counts before and after temporal quality control were compared using paired tests: normality was checked with the Shapiro-Wilk test and since the data were normally distributed, a paired t -test with within-subjects effect size was used. Percent rejection was reported as the average percentage of components removed across sessions. Signal-to-noise improvement was computed within each session as the mean ratio of denoised to raw traces, expressed in decibels, and compared using the same paired testing framework. Noise power was estimated from the standard deviation of the trace derivative, while signal power was taken as the variance of the trace. Decay times were derived from the autoregressive parameters and summarized with mean, median, and range.

Spatial refinement was evaluated per session and summarized across sessions. Component counts before and after refinement were compared with the same paired testing framework as above. Spatial organization was quantified with nearest-neighbor distances of component centroids, calculated using KD-trees, and summarized by the session-level mean and median. Density maps were generated to visualize local packing of components. Demixing quality was assessed by comparing temporal correlations among spatially overlapping neighbors before and after refinement. Morphology was summarized with measures such as compactness (perimeter squared divided by area) and centroid jitter across iterations; pre- and post-refinement values were compared per session with paired tests and effect sizes. When manual masks were available, specificity was quantified with intersection-over-union scores, reported as median and interquartile range per session.

2.4.16 Citations

2.5 Acknowledgments

Funding for this work and for the authors' research effort was provided by the National Institute on Drug Abuse (T32DA007278, R01DA052618, and R00DA052571) and by the Puget Sound VA Research and Development Pilot Grant.

Chapter 3

Astrometric Tracking of Cells Across Imaging Sessions

Submitted as: Peden-Asarch A, Honan L, Quinn J, Coffey KR, Neumaier JF. Stars2Cells: Astrometric Tracking of Cells Across Imaging Sessions. Preprint on bioRxiv.

3.1 Introduction

Chronic two-photon microscopy through cranial windows first established that the same cortical neurons could be revisited over days and weeks, enabling imaging-based single-cell investigations of learning, plasticity, and representational drift *in vivo* (Trachtenberg et al., 2002; Holtmaat et al., 2009; Driscoll et al., 2017). Miniaturized one-photon fluorescence microscopes (“miniscopes”) extended this access to freely behaving animals and to deep-brain structures inaccessible through head-fixed objectives (Ghosh et al., 2011a; Ziv et al., 2013a; Aharoni and Hoogland, 2019a; Ziv and Ghosh, 2015a; Zhou et al., 2018a). Regardless, re-identifying a given neuron from one session to the next is the prerequisite for any longitudinal investigation, and the difficulty in reliably identifying the same neurons from session to session has prevented researchers from taking full advantage of recurrent imaging.

To quantify this disparity, we surveyed PubMed, where two-photon and miniscope publications have grown steeply over the past decade (Fig. 3.1a). Combined, ~8,484 unique last authors have published two-photon calcium-imaging work (with a further ~500 UCLA Miniscope and ~600 Inscopix laboratories estimated from manufacturer/community figures; Fig. 3.1b), yet only ~220 (~2.3%) unique last authors have published longitudinal cross-session imaging since 2021. Tool-level adoption is still sparser with citation counts of each tool’s reference publication (OpenAlex, accessed 06/2026) as follows: CellReg (Sheintuch et al., 2017b) leads with 435 citations, reflecting both an eight-year head start and the footprint-correlation family’s incumbency in 1P miniscope data. SCOUT (Johnston et al., 2022) follows with 12 citations and “Track2p” (Majnik et al., 2025) with 1

citation, spanning longitudinal-pipeline and developmental-2P use, respectively; CaliAli (Vergara et al., 2025a) has 4 citations; both were published within the past year, and thus still below the citation horizon. Among cross-session image-registration tools, PatchWarp (Hattori and Komiyama, 2022) has 29 citations and the fully affinity-invariant FOV-alignment method FAIM (Li et al., 2020) has 3 citations; and EMC² (Lagache et al., 2021), for continuous tracking within the session, has 33 citations.

These tools span three layers of the pipeline, distinguished by the input each consumes (Fig. 3.1c): *movie-level* methods (CaliAli (Vergara et al., 2025a)) register raw video on landmarks such as vasculature before extraction; *footprint-level* methods (CellReg (Sheintuch et al., 2017b), CaImAn’s register_multisession (Giovannucci et al., 2019b), SCOUT (Johnston et al., 2022)) match per-session ROI masks by spatial correlation; and *centroid-level* matching, where S2C operates, recovers correspondence directly from the 2D point geometry any extractor produces. Each layer trades one cue for another (Fig. 3.1d): movie-level needs visible vasculature, footprint-level needs ROI shape stability across days, and centroid-level methods discard both but runs on any extractor’s output across any modality (1P, 2P, mesoscope, hand-drawn), handles rotation by construction, and applies to archived data. S2C delivers this as a GUI-driven standalone application for macOS and Windows (Fig. 3.1e) that carries raw recordings through to multi-session neuron tracks end-to-end.

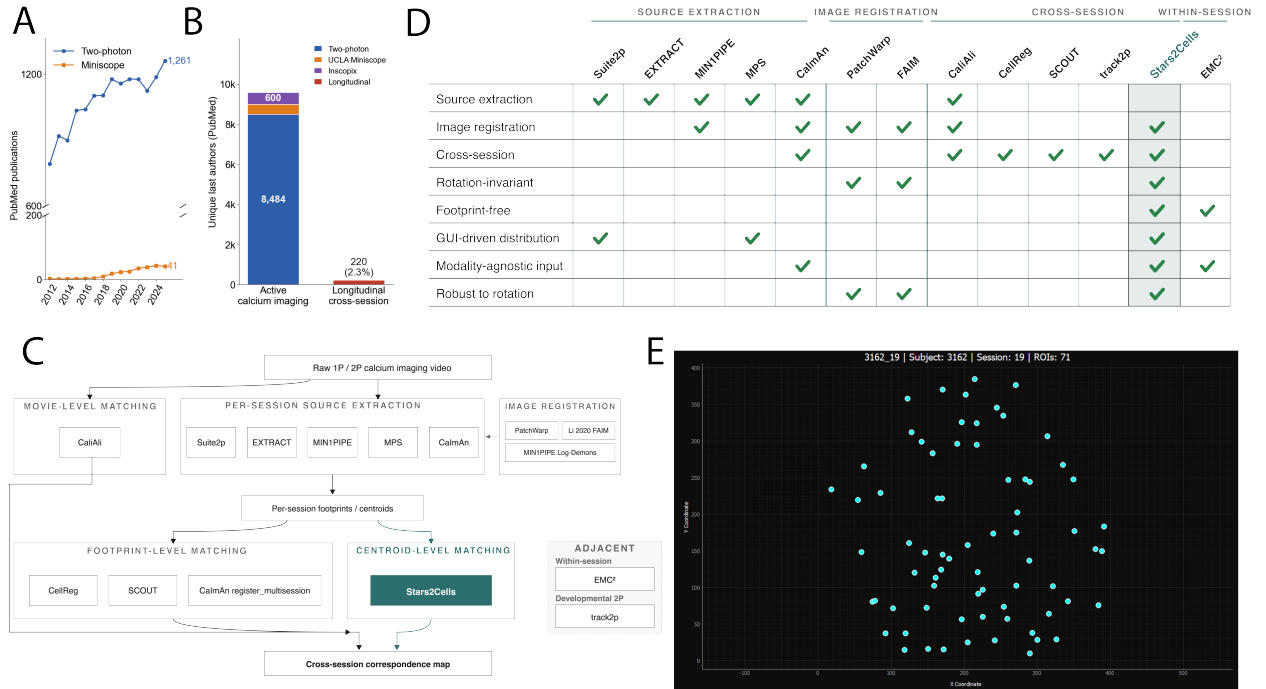


Figure 3.1: The longitudinal tracking gap, existing tool landscape, and the GUI-driven S2C application. **a**, PubMed publications per year for “two-photon” (blue) and “miniscope” (orange) (2012–2025), showing rapid growth in active calcium imaging. **b**, Active imaging labs versus labs publishing longitudinal cross-session tracking. **c**, Feature comparison of the calcium-imaging analysis tool landscape. **d**, Layer diagram of the cross-session analysis pipeline. **e**, Stars2Cells GUI-driven standalone application.

Two converging bottlenecks have kept this longitudinal imaging out of the scientific zeitgeist. First, raw single-photon miniscope video processing historically required deep computational expertise. Pipelines such as CaImAn and Minian asked end users to assemble Python or MATLAB environments, configure GPU/parallel backends, manually tune CNMF/CNMF-E parameters, and inspect results from the command line, which alone can take new groups weeks to months to operationalize (Giovannucci et al., 2019b; Dong et al., 2022a). MPS (Peden-Asarch et al., 2026c) addresses this upstream side by consolidating preprocessing, motion correction, source extraction, deconvolution, and quality control behind a single graphical interface, removing the engineering overhead that historically gated entry for non-engineering groups. Second, reliably matching the same cells across recording days has remained unsolved at the level of inter-session FOV variability that single-photon recordings regularly produce. The incumbent approach matches cells by correlating their per-session ROI footprints (hereafter *ROI matching*), the strategy behind CellReg (Sheintuch et al., 2017b) and the footprint-correlation methods that followed it. It assumes tightly co-registered fields of view and stable footprint shape across days, both of which break under the baseplate-reattachment rotations, translational drift, and detection dropout that chronic single-photon recordings routinely produce. However, this disparity is not driven by lack of interest in longitudinal questions, but reflects the practical difficulty of reliably matching neurons across sessions with existing tools, layered on top of the overhead of installing, configuring, and operating Python-based pipelines.

The connection to astronomy is the idea that first prompted this project: how does a telescope point to an unfamiliar patch of sky and know where in space it is? The answer is astrometric plate-solving, which is the routine, well-validated procedure by which astrometric pipelines such as Astropy and Astrometry.net match an unannotated star field to a catalog despite unknown rigid transformations (Hogg et al., 2008; Lang et al., 2010). The solution emerges from local geometry where each star is described by the geometry of its four-star neighborhood (a quad), and the resulting normalized descriptor is invariant to rotation, translation, and uniform scaling (Hogg et al., 2008; Lang et al., 2010). Matching is performed in descriptor space, followed by geometric verification via RANSAC, yielding robust correspondences even when large fractions of the field are obscured. However, this machinery only solves *where* a telescope is pointing, and thus, does not assign astronomical identities to individual stars. Therefore, S2C is the natural follow-up intended for cross session the neuron-tracking.

The centroid coordinates produced by MPS (Peden-Asarch et al., 2026c), Suite2p (Pachitariu et al., 2017), CaImAn (Giovannucci et al., 2019b), Minian (Dong et al., 2022a), MIN1PIPE (Lu et al., 2018), EXTRACT (Inan et al., 2021a), or hand-drawn ROIs all naturally feed into S2C. We designed MPS and S2C to form a complete GUI-driven workflow from raw miniscope video to longitudinal neuron tracking, distributed as GUI-driven standalone applications for both macOS and Windows (Fig. 3.1e). The pipeline proceeds in five automated steps: (1) quad descriptor generation from centroid constellations using a diagonal-first algorithm with KNN local and random long-range sampling; (2) descriptor-space matching with an automatically calibrated similarity threshold that scales as $\sqrt{N}$, where N is the number of neurons; (3) RANSAC geometric verification

to reject false-positive descriptor matches; (4) Hungarian assignment to produce globally optimal one-to-one neuron correspondences; and (5) consolidation of pairwise matches into multi-session neuron tracks. Each step is visualized and inspectable through an integrated GUI; each build bundles a self-contained Python runtime and all dependencies (a signed and notarized DMG for macOS, a portable executable for Windows), so no manual dependency configuration is required.

The dorsomedial striatum (DMS) is the canonical associative striatal hub for action-outcome learning, value computation, and goal-directed control (Yin and Knowlton, 2006; Balleine et al., 2021), and dorsal/dorsomedial striatum is causally implicated in relapse to opioid seeking after voluntary abstinence (Ma et al., 2025), with dorsal-striatal connectivity tracking the incubation of opioid craving (Fredriksson et al., 2021a). DMS dopamine signaling longitudinally tracks the emergence of compulsive reward seeking across self-administration (Seiler et al., 2022). Yet despite this causal centrality, to our knowledge every published *in vivo* recording in dorsomedial striatum *cell bodies* during *opioid* self-administration to date has been bulk neurochemical (fiber photometry) via dopamine, glutamate, or aggregated Ca^{2+} activity (Seiler et al., 2022; Legaria et al., 2022; Coffey et al., 2023; Bender et al., 2024) leaving the cellular composition and longitudinal evolution of DMS dynamics during opioid taking uncharacterized at single-cell resolution. How DMS output neurons reorganize across chronic opioid exposure has been characterized almost exclusively *ex vivo*, and *in vivo* only in nucleus accumbens via miniscopes (Graziane et al., 2016; Hearing et al., 2016; Engeln et al., 2020; Domingues et al., 2025). In parallel, behavioral-economics demand procedures (Hursh and Silberberg, 2008; Bentzley et al., 2014) yield scale-invariant, orthogonal measures of intake (Q_0) and motivation (α) that uniquely predict compulsivity, reinstatement, and pharmacotherapy efficacy, including for fentanyl (Fragale et al., 2019; McConnell et al., 2021; Coffey et al., 2023). Yet these measures have been combined only with bulk neurochemical signals and not with single-cell activity. Thus, since population-level signals are known to mask near-complete ensemble turnover across hippocampus, cortex, and striatum, then asking whether the same DMS neurons carry the press response across the demand curve is impossible without same-cell tracking. The MPS→S2C stack applied here provides the first cell-resolved longitudinal recording of individually tracked dorsomedial striatum neurons across an escalating opioid demand curve, to our knowledge.

The website for download is available at <https://ariasarch.github.io/Stars2Cells/>, a sample walkthrough at https://github.com/ariasarch/Stars2Cells_Sample_Code, and the source code at https://github.com/ariasarch/Stars2Cells_GUI.

3.2 Results

3.2.1 Quad descriptor pairing from neuron centroid constellations

S2C represents each neuron’s spatial context by enumerating four-neuron patterns (quads) from the centroid field and computing a normalized geometric descriptor for each (Fig. 3.2a). Given N centroid positions in a session, the pipeline constructs candidate diagonals by combining k -nearest-neighbor local edges with random long-range edges, ensuring descriptors capture both fine local structure and FOV-spanning geometry (Fig. 3.2b). For each diagonal, perpendicular heights are computed for all other neurons, and the top- K third points by height are retained (default $K = 25$ –tunable). Third points are then paired to form quads, and a four-dimensional descriptor (x_C, y_C, x_D, y_D) is computed by projecting the two non-diagonal vertices onto the coordinate frame defined by the quad’s longest pairwise edge, normalizing by the diagonal length.

A key algorithmic innovation is the ownership guard: each quad is emitted exactly once, from the single diagonal equal to its longest pairwise edge (Methods; Fig. 3.2b). This avoids both the intractable $\binom{N}{4}$ scaling of brute-force enumeration and the $6\times$ duplicate emission of a guard-free construction.

Quad saturation analysis confirmed that the median nearest-neighbor distance in descriptor space ($NN_d = 0.0101$) exceeded the median descriptor blur ($\sigma_d = 0.0041$) for all sessions, with a saturation ratio of 2.46, indicating that the descriptor field was not saturated and individual neurons remained distinguishable.

Accurate neuron matching requires a larger similarity threshold when neuron count grows. Thus, this is built into the next step of S2C via descriptor matching that requires a similarity threshold τ below which two quad descriptors are considered a match. As a field densifies, local-diagonal lengths shrink and the per-coordinate blur of a matched pair rises. Thus, since same-cell descriptor blur grows with neuron count, a fixed threshold rejects true matches at large N and admits false ones at small N , so τ must grow with field size. The empirical $\sqrt{N}$ scaling is illustrated at the four benchmark neuron counts in Fig. 3.2c. S2C automates the per-animal calibration of C in Step 1.5 by sweeping thresholds on sampled quad subsets for each animal and fitting via least-squares regression through the origin. The resulting $\tau = C\sqrt{N}$ is applied per animal as the Step 2 descriptor threshold; it is a geometry-motivated estimate rather than a precisely calibrated value (Supplementary Note 2), with any residual miscalibration absorbed downstream by the consistency filter and RANSAC, so that registration accuracy is robust to its exact value (98.4% mean F1; Fig. 3.4).

Nonetheless, raw descriptor matches still contain both true correspondences and false positives from coincidental descriptor similarity. S2C applies a two-stage filtering process. First, a consistency filter removes matches whose implied spatial transformation is inconsistent with the consensus. Second, vectorized RANSAC (Fischler and Bolles, 1981) estimates the dominant rigid transformation between session pairs from matched quad centroid positions, retaining only geometrically consistent inlier matches (Fig. 3.2d).

The surviving quad-level matches are aggregated into a neuron-level vote matrix V , where $V[i, j]$ counts how many verified quad matches jointly contain reference neuron i and target neuron j . The Hungarian algorithm (Kuhn, 1955) is applied to the cost matrix $(\max(V) - V)$ to produce globally optimal one-to-one neuron assignments. A cost-threshold sweep identifies the assignment threshold that maximizes matched neurons while preserving geometric coherence (Fig. 3.2e). A two-pass solve recovers correspondences that a single conservative cutoff would discard: Pass 1 enforces a tight $1.0\times$ RANSAC-residual cutoff (tunable), Pass 2 re-solves a smaller Hungarian over the unmatched residual under a relaxed $2.0\times$ cutoff, and matches with residual exceeding $2.0\times$ the RANSAC threshold (tunable) are rejected outright (Fig. 3.2f).

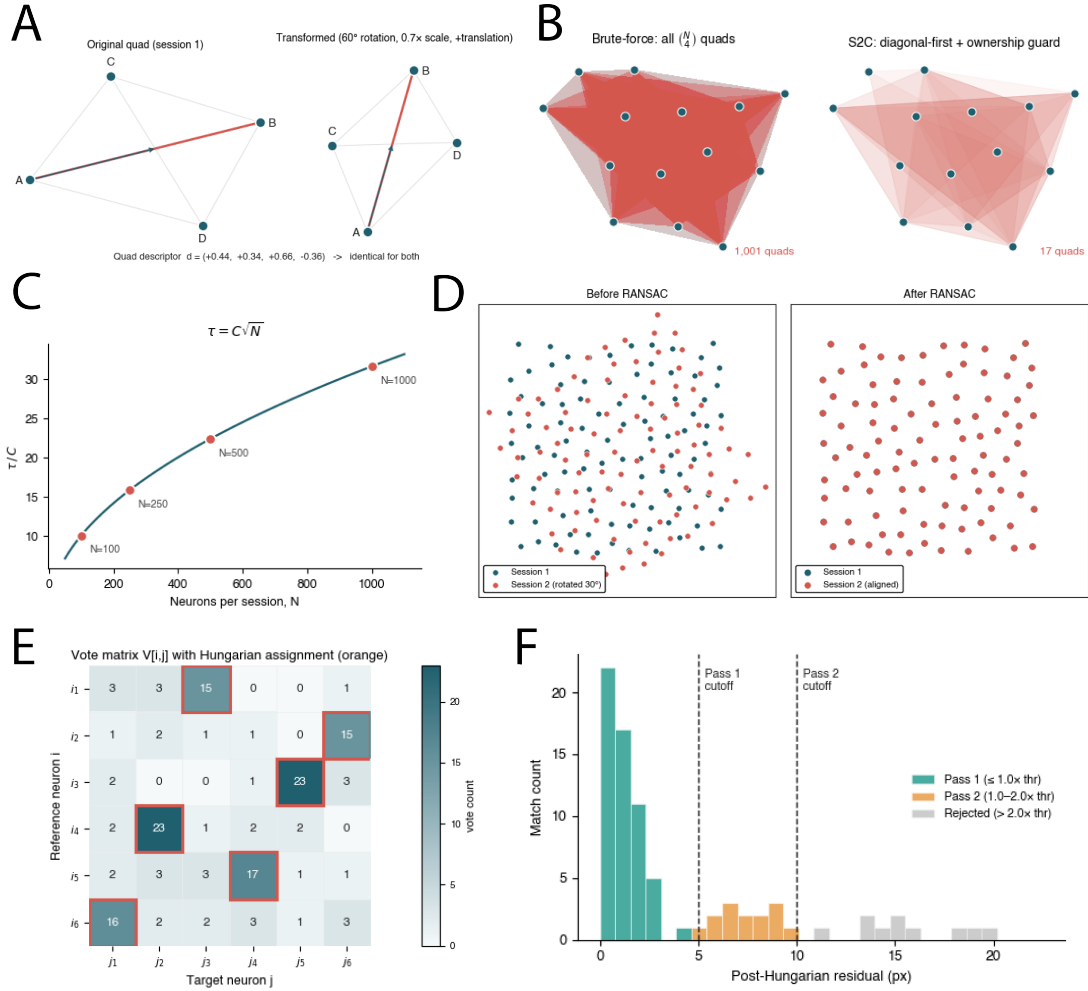


Figure 3.2: The Stars2Cells pipeline. **a**, Astrometric metaphor: a four-star quad’s relative geometry, normalized on its longest edge. **b**, Step 1, diagonal-first quad enumeration with ownership guard. Each translucent polygon spans the four neurons of one enumerated quad; cumulative compositing renders quad density as color saturation. *Left*: brute-force enumeration of all $\binom{N}{4}$ quads on an example field ($N = 14$ neurons; 1,001 quads) saturates into an unreadable red blob where no single quad is resolvable. *Right*: S2C’s diagonal-first construction ($k_{\text{local}} = 15$ KNN local diagonals plus k_{random} long-range partners per neuron at production scale; scaled down to $k_{\text{local}} = 4$, $k_{\text{random}} = 2$ here so the $N = 14$ stays readable) emits each unique quad exactly once from its longest pairwise edge, yielding a sparse selective set in which individual quads remain visible. Complexity: $O(NkK^2)$ vs. $O(N^4)$ for brute force. **c**, Step 1.5, empirical similarity-threshold scaling: the optimal threshold τ scales as $\tau = C\sqrt{N}$ across the four benchmark neuron counts ($N = 100, 250, 500, 1,000$). **d**, Step 2.5, vectorized RANSAC: 100 simulated neurons (Poisson-disk-sampled in a 600×600 FOV) with session 2 rotated 30° about the FOV centre plus a small translation (left); after RANSAC fits the rigid transform from $K = 1,000$ batched-SVD candidate transforms and the inverse is applied, session 2 (orange) registers exactly to session 1 (teal) (right). **e**, Step 3, vote matrix and Hungarian assignment: each verified quad votes for $4 \times 4 = 16$ neuron pairings; degree-normalized votes plus a spatial-residual term form the cost matrix, which is padded with rectangular dummy slots so unmatched neurons can stay unmatched. **f**, Pass 1 versus Pass 2 recovery (residual distribution by pass): Pass 1 enforces a tight $1.0 \times$ RANSAC residual cutoff (green); Pass 2 re-solves a smaller Hungarian over the unmatched residual under a relaxed $2.0 \times$ cutoff (orange), recovering correspondences that Pass 1 conservatively rejected; matches with residual $> 2.0 \times$ the RANSAC threshold are rejected outright (grey).

3.2.2 Activity correlation alone cannot identify neurons across sessions

A natural question is whether activity-based features such as cross-session calcium-trace correlation could substitute for or complement geometric matching. We bound its contribution from above with a deliberately aggressive idealized scenario that hands the temporal-correlation matcher every advantage geometry could provide: 1,000 simulated neurons placed at *identical* spatial positions across five sessions. Because positions are identical, any geometric or footprint-based matcher scores perfectly, so the test isolates the contribution of activity alone.

For each pair of sessions, we computed the full $1,000 \times 1,000$ Pearson correlation matrix between z -scored traces (in float32 row-chunks of 100 to control memory) and assigned each neuron in session s_1 to the maximally correlated neuron in session s_2 . Performance was evaluated against ground-truth identity across recording durations of 10, 20, 60, and 180 min (6,000 to 108,000 frames per session). At every tested duration, mean matching accuracy across the 10 session pairs fell within the chance band: 0.12% [0.06, 0.19] at 10 min ($p = 0.53$ vs. chance), 0.14% [0.09, 0.20] at 20 min ($p = 0.38$), 0.07% [0.01, 0.14] at 60 min ($p = 0.56$), and 0.15% [0.08, 0.24] at 180 min ($p = 0.53$); chance is $1/N = 0.10\%$ (Fig. 3.3). Thus, the activity-only matcher recovered no meaningful identity signal as each fell within the chance band at every duration.

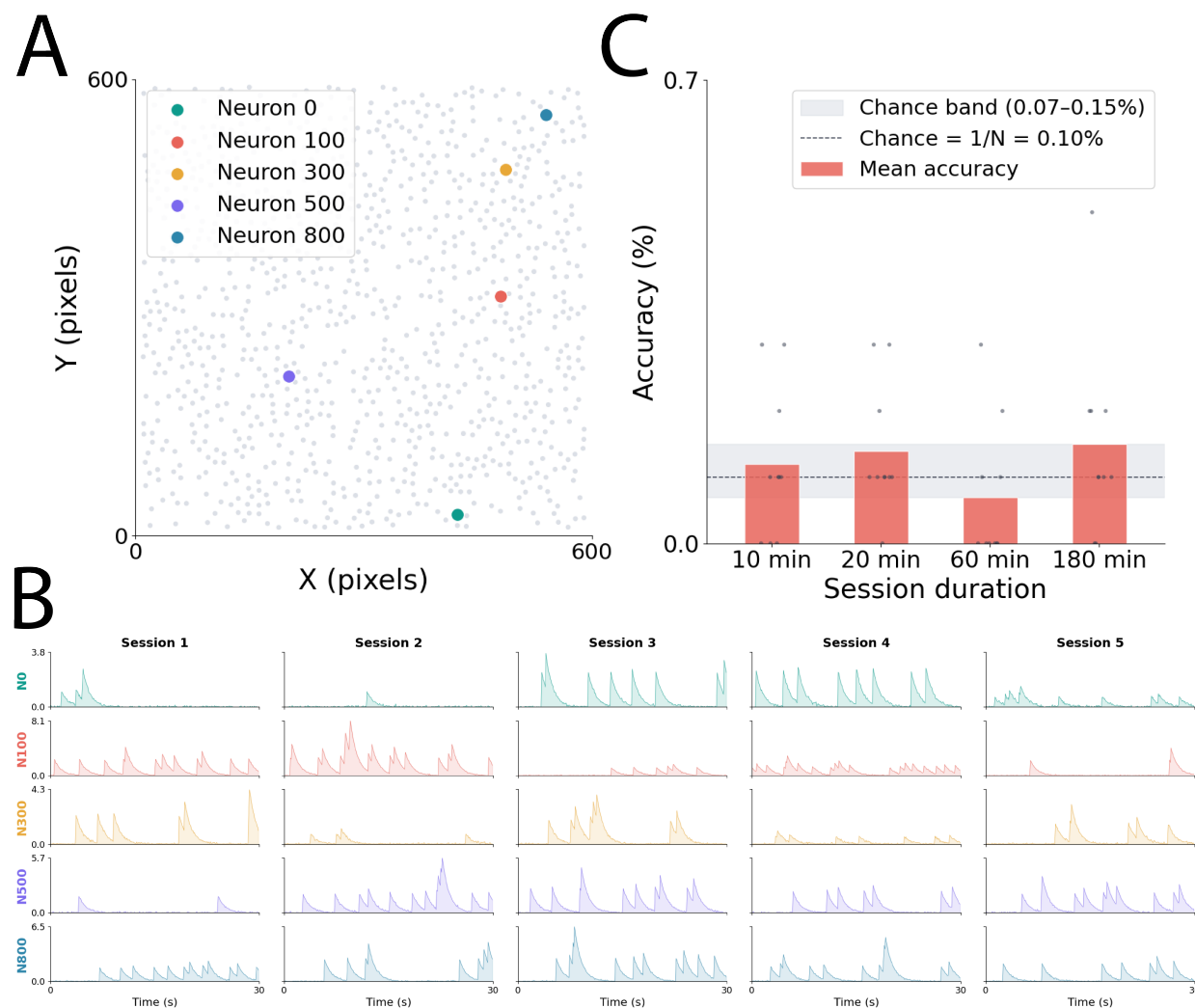


Figure 3.3: Temporal correlation alone is chance-level for cross-session matching. **a**, Spatial map: 1,000 simulated neurons placed by Poisson-disk sampling in a 600×600 px FOV; the same layout is reused across all 5 sessions, so a geometric matcher always would score 100%. Five example neurons are highlighted (colored) for panel **b**. **b**, Example AR(1) calcium traces (first 30 s shown) for the five highlighted neurons across sessions 1–5. Despite identical spatial identity, dynamics are statistically independent across sessions – each session reflects a fresh draw from that neuron’s generative distribution. **c**, Mean cross-session matching accuracy (mean $\pm$ scatter across 10 session pairs) from a max-correlation matcher remains within the chance band (0.07–0.15%; chance $1/N = 0.10\%$, dashed line) at all four tested durations (10, 20, 60, 180 min); $p > 0.3$ vs. chance at every duration.

3.2.3 Synthetic validation across recording modalities

We validated S2C on a synthetic benchmark of 8 inter-session degradation conditions plus an identity floor evaluated at four neuron-count tiers (100, 250, 500, 1,000), chosen to span the densities seen across calcium-imaging preparations: ~ 50 – 200 neurons per FOV in deep-brain miniscope recordings of DMS, habenula, dorsal raphe, and BNST (Zhou et al., 2018a); ~ 200 – 500 in cortical and hippocampal miniscope recordings (Ziv et al., 2013a; Cai et al., 2016b); and ~ 500 – $1,000+$ in two-photon cortical imaging (Pachitariu et al., 2017). Degradation parameters were selected to bracket the inter-session perturbations we and others have observed in chronic miniscope recordings, while also extending into scenarios that the field’s hardware and surgical practice could plausibly produce: baseplate-reattachment rotation and animal head motion against the implant (uncorrected by MPS’s translational-only motion correction), residual translational drift downstream of MPS, ROI dropout from CNMF detection variability, and centroid localization noise. The ranges therefore combine direct empirical observation with a forward-looking margin for the more challenging operating conditions chronic experiments may encounter.

S2C is accurate and robust across the full benchmark parameter space. S2C achieved a mean F1 score of 98.4% across 1,262 paired runs spanning 8 conditions and 32 productive (condition $\times$ neuron-count) cells, with precision 99.9% and recall 97.2% (Fig. 3.4c). Performance was robust across neuron counts: F1 = 99.7% at $N = 100$, 99.9% at $N = 250$, 97.0% at $N = 500$, and 97.3% at $N = 1,000$ (Fig. 3.4d). Across tiers (illustrated at $N = 250$), S2C reached 99.9% F1 on Tier A (2P-like), 99.9% on Tier B (Miniscope-Like), and 99.9% on Tier C (Extreme) (Fig. 3.4e). Per-tier dips at higher neuron counts were driven by recall (precision remained near-saturated, $\geq 99.4\%$ across all neuron counts), reflecting genuinely missed correspondences in the densest, most-perturbed conditions rather than spurious matches. Transitivity consistency across five chained sessions was high: pairwise F1 across the four consecutive transitions remained essentially flat at 98.3%, 98.4%, 98.4%, and 98.4% for $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow D$, and $D \rightarrow E$ respectively, indicating that registration accuracy is independent of position within the chain and does not degrade as sessions accumulate (Fig. 3.4f).

S2C achieved higher F1 than ROI-based matching at every neuron count and on every condition tested. Pooling within each neuron-count tier under the worst-case footprint scenario (complete shape perturbation; Fig. 3.4g): at $N = 100$, S2C reached F1 = 99.7% vs. ROI-based matching’s 45.6% ($\Delta F1 = +54.1$ pp, 95% CI [+50.0, +58.1], paired Cohen’s $d_z = +1.49$, $p = 1.1 \times 10^{-51}$ Holm-corrected); at $N = 250$, F1 = 99.9% vs. 39.6% ($\Delta = +60.3$ pp [+56.2, +64.3], $d_z = +1.67$, $p = 1.8 \times 10^{-53}$); at $N = 500$, F1 = 97.0% vs. 33.1% ($\Delta = +63.9$ pp [+60.3, +67.5], $d_z = +1.98$, $p = 3.3 \times 10^{-52}$); at $N = 1,000$, F1 = 97.3% vs. 25.9% ($\Delta = +71.4$ pp [+67.9, +74.6], $d_z = +2.34$, $p = 3.5 \times 10^{-53}$). Pooled across all 1,262 paired runs, S2C reached F1 = 98.4% vs. ROI-based matching’s 36.0% (paired mean $\Delta F1 = +62.4$ pp [+60.5, +64.4], $d_z = +1.81$, $p < 0.001$; Fig. 3.4g, dumbbell panels per neuron count). Restricting the pool to the empirically-grounded conditions (Tiers A and B) demonstrates that the pooled gap is not an artifact of the most extreme conditions: S2C F1 = 98.4% vs. ROI-based matching 38.7% (paired mean $\Delta F1 = +59.7$ pp [95%

CI +57.6, +61.8], $d_z = +1.67$, $n = 1,102$, $p < 0.001$). At every individual condition tested, the F1 advantage was statistically significant after Holm correction (all $p < 0.005$), and S2C’s precision was uniformly $\geq 97.7\%$ across all conditions (pooled precision $\geq 99.4\%$ at every neuron count).

The ROI-based matching deficit is geometric, not footprint-driven. Giving ROI matching a perfectly stable per-neuron footprint shape did not improve its registration accuracy. Pooled across all paired runs, the best-case scenario (footprint shape fixed and unique per neuron, reused unchanged across sessions) reached ROI-matching F1 = 36.2%, essentially identical to the worst-case scenario’s 36.0%, and the S2C advantage was essentially identical between scenarios (best-case grand-pooled $\Delta F1 = +62.3$ pp [95% CI +60.4, +64.2], $d_z = +1.81$, $n = 1,262$; worst-case = +62.4 pp [+60.5, +64.4], $d_z = +1.81$, $n = 1,262$). The two scenarios tracked each other at every neuron-count tier, with ROI F1 differing by ≤ 0.4 pp (45.9% vs. 45.6% at $N = 100$, 39.6% vs. 39.6% at 250, 33.1% vs. 33.1% at 500, and 26.3% vs. 25.9% at 1,000). Across individual (condition $\times$ neuron-count) cells the two scenarios differed by at most 4.3 pp with no consistent direction indicating that footprint shape jitter is downstream noise rather than a decisive identity signal. ROI matching’s failure mode on this benchmark traces to the *geometric* perturbations (rotation, translation, dropout) breaking the spatial-correlation signal upstream of any benefit footprint-shape stability could confer. The pattern is most stark at high density: at 1,000 neurons, where the mixture model collapses across the Tier-B and Tier-C stress conditions (Per-condition, the five highest-displacement conditions are B_rot $\Delta = +95.3$ pp, B_combined $\Delta = +93.9$ pp, B_drift $\Delta = +92.1$ pp, C_walk $\Delta = +87.9$ pp, and B_dropout $\Delta = +76.1$ pp, while the three milder Tier-A conditions span $\Delta = +10.6$ to +61.5 pp; all 32 (condition $\times$ neuron-count) detailed cell outputs are in Supplementary Table B.1). We additionally ran both pipelines on CellReg’s publicly released sample sessions, but those sessions carry no ground-truth identity, so any comparison reduces to raw match counts that cannot be scored for correctness; we therefore confine the quantitative precision/recall comparison to the synthetic benchmark.

S2C transfers to two-photon data acquired by another laboratory. Run blind on 431 Allen Brain Observatory Visual Coding 2P containers S2C reproduced the Allen Institute’s own cross-session registration with median F1 = 0.97 (IQR 0.46–1.00) for A→B and 0.97 (0.47–1.00) for B→C, confirming the method generalizes beyond one-photon miniscope data and our own recordings (Methods, Section 3.4.8).

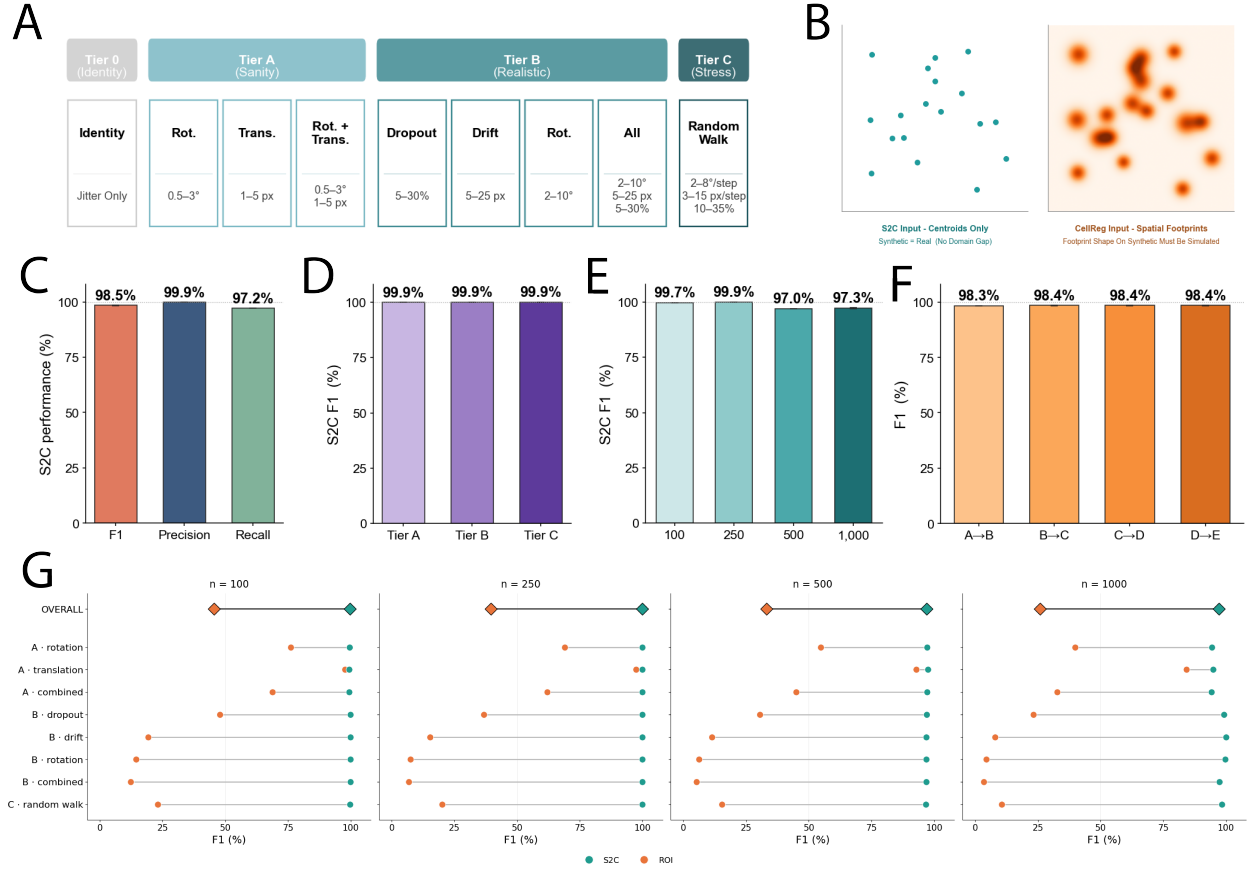


Figure 3.4: S2C is robust across the full benchmark parameter space and outperforms ROI-based matching at every condition. **a**, Benchmark structure: 8 productive conditions plus 1 identity sanity-floor (Tier 0 identity / Tier A (2P-Like) 3 / Tier B (Miniscope-Like) 4 / Tier C (Extreme) 1) $\times$ 4 neuron counts (100, 250, 500, 1,000) $\times$ 8 animals $\times$ 5 seeds $\times$ 5 consecutive sessions. **b**, Domain-gap-free validation: because S2C operates exclusively on 2D centroid coordinates, the input distribution is identical for synthetic and real recordings. ROI-based matching, in contrast, requires footprint shape that must be simulated approximately. **c**, Overall F1, precision, and recall pooled across all conditions and neuron counts. **d**, F1 stratified by neuron count. **e**, F1 stratified by tier at fixed $N = 250$. **f**, Per-transition F1 across five-session chains: pairwise F1 for the four consecutive transitions (A→B, B→C, C→D, D→E). Performance is stable across chain position, with no per-step degradation. **g**, S2C versus ROI-based matching: dumbbell plot of paired F1 (S2C, teal ●; ROI-based matching, orange ●) under the worst-case (perturbed) footprint scenario, one column per neuron count (100, 250, 500, 1,000). Each row is one of 8 degradation conditions (Tier-A rotation/translation/combined; Tier-B dropout/drift/rotation/combined; Tier-C random walk); diamond markers (**OVERALL**) pool all 8 conditions per neuron count (up to $n = 320$ paired runs each). S2C outperforms at every cell, with a pooled S2C – ROI-based matching gap of +54.1, +60.3, +63.9, and +71.4 pp at $N = 100, 250, 500, 1,000$. The ROI-based matching deficit is geometric (rotation/translation/dropout breaking the spatial-correlation signal upstream) rather than footprint-shape-driven: per-neuron-count $\Delta F1$ is essentially identical between best-case and worst-case conditions (matched-condition surface, 100–500 neurons, all 8 conditions; $|\Delta_{\text{best-case}} - \Delta_{\text{worst-case}}| \leq 0.3$ pp). Grand-pooled paired mean $\Delta F1 = +62.4$ pp [+60.5, +64.4], $d_z = +1.81$, $p < 0.001$ across 1,262 paired runs (per-condition: paired Wilcoxon signed-rank, Holm-corrected; effect sizes paired Cohen’s d_z ; 95% CIs bias-corrected bootstrap).

3.2.4 Longitudinal DMS tracking across a fentanyl behavioral economics paradigm

To evaluate the full MPS→S2C stack on a real chronic miniscope dataset, we applied the pipeline to dorsomedial striatum (DMS) calcium imaging in Long-Evans rats ($n = 3$ male, $n = 3$ female; Methods).

Raw single-photon miniscope videos were processed through MPS (Peden-Asarch et al., 2026c), yielding 119 ± 27 segmented neurons per session per animal (median 111, range: 79–183; $n = 24$ sessions across 6 animals). Viral targeting of the DMS is shown in Fig. 3.5a. For a representative session, the same field of view is shown as a temporal mean projection after MPS preprocessing and window cropping (Fig. 3.5b) and the resulting MPS-extracted spatial footprint map (Fig. 3.5c). Event-aligned, baseline-subtracted activity traces for the most active segmented neurons in that session are shown in Fig. 3.5d. Lastly, S2C registered these footprints across sessions, identifying neurons recorded on different days: for a representative animal, cells tracked between two sessions (222 and 70 $\mu\text{g/mL}$) are shown in green, linked across the two fields of view (Fig. 3.5e).

Across the six animals, S2C tracked DMS neurons through the four-session registration chain ($222.0 \rightarrow 70.0 \rightarrow 22.0 \rightarrow 0$ $\mu\text{g/mL}$). At each consecutive step, S2C matched a much larger fraction of neurons than ROI matching: $69.1\% \pm 7.2\%$ vs. $13.0\% \pm 3.3\%$ ($222 \rightarrow 70$), $73.6\% \pm 4.4\%$ vs. $17.6\% \pm 5.7\%$ ($70 \rightarrow 22$), and $77.4\% \pm 2.0\%$ vs. $19.4\% \pm 4.0\%$ ($22 \rightarrow 0$) (mean $\pm$ SEM, $n = 6$; Fig. 3.5f). Because S2C registers sessions sequentially, this per-step advantage compounds: end-to-end, S2C retained $41.6\% \pm 3.4\%$ of first-session (222 $\mu\text{g/mL}$) neurons across the full $222 \rightarrow 0$ chain versus $13.0\% \pm 3.9\%$ for ROI matching, representing a 3.2-fold improvement in longitudinal yield. However, since GCaMP-based detection is activity-dependent, a neuron that falls silent in any intervening session cannot be segmented and drops from the sequential chain. Thus, given that per-session yield was stable across the imaging window (Methods), this 41.6% is best read as a conservative lower bound on true same-cell persistence rather than a measure of expression or indicator loss.

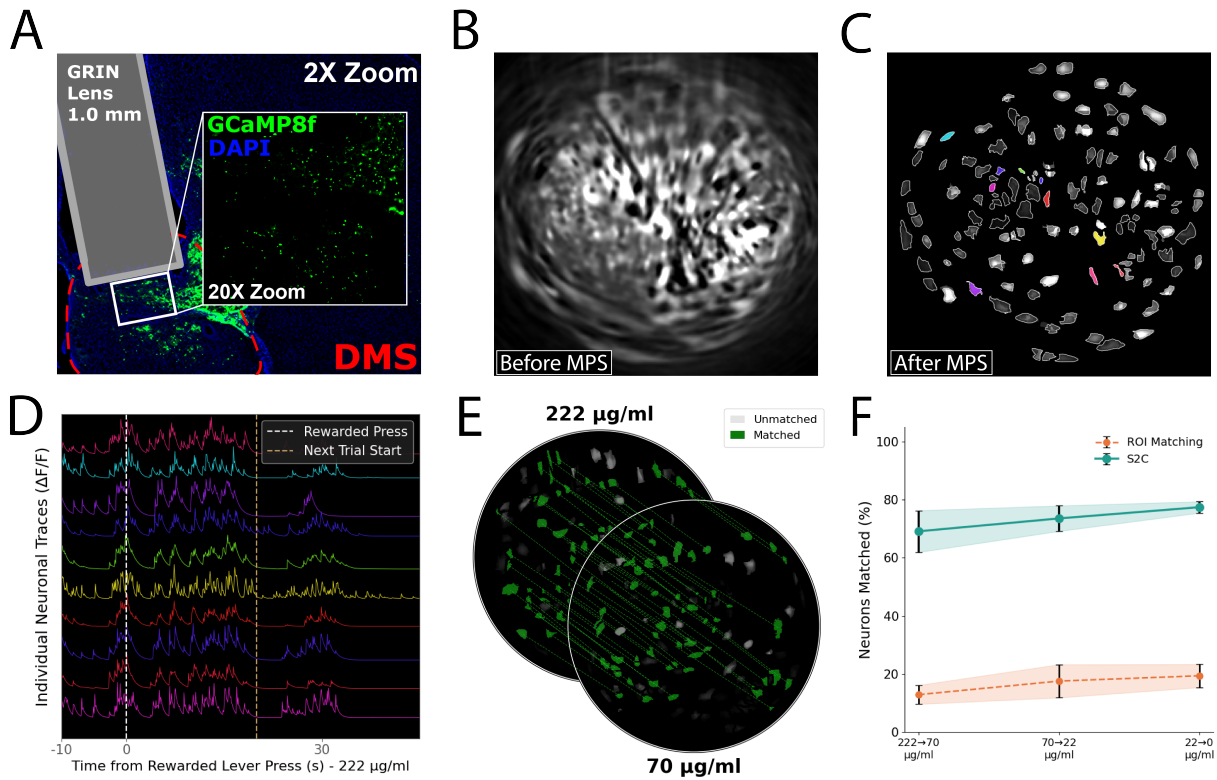


Figure 3.5: End-to-end application of S2C via same-cell longitudinal tracking of DMS neurons across an escalating fentanyl demand curve. **a**, Representative histology confirming GRIN-lens placement and GCaMP8f expression in dorsomedial striatum (DMS). Coronal section showing GCaMP8f (green) with DAPI nuclear counterstain (blue); the DMS boundary is outlined (red dashed) and the 1.0 mm GRIN-lens tract overlaid (2 $\times$). Inset (20 $\times$): high-magnification view of GCaMP8f-expressing DMS neurons beneath the lens face. **b**, Temporal mean projection of the dorsomedial striatum field of view after MPS preprocessing (degrow, denoising, and motion correction with frame removal) and window cropping. **c**, Spatial footprint map extracted from the same field of view by MPS. **d**, Baseline-subtracted activity traces for the ten most active MPS-segmented DMS neurons over a window spanning the pre-event baseline through the post-event period, aligned to a representative event (white dashed line); each trace is colored to match its footprint in **c**. **e**, Spatial footprint maps from two example sessions (top, 222 $\mu\text{g/mL}$; bottom, 70 $\mu\text{g/mL}$) for an example animal, shown as stacked, circle-cropped fields of view. Neurons matched across the two sessions are filled in green and linked by dashed lines and unmatched neurons are shown in gray. **f**, Per-step registration performance: fraction of neurons matched across each consecutive dose transition (222 $\rightarrow$ 70, 70 $\rightarrow$ 22, 22 $\rightarrow$ 0 $\mu\text{g/mL}$) for Stars2Cells (S2C) versus ROI matching (mean $\pm$ SEM, $n = 6$).

3.2.5 Behavioral demand and longitudinal DMS population dynamics

Animals performed oral fentanyl self-administration under a behavioral-economics demand-curve schedule (Fig. 3.6a, b). The resulting demand curves followed the canonical exponential decay shape, with population-level Q_0 and α parameters consistent with published rat oral fentanyl demand work (Coffey et al., 2023) (Fig. 3.6c; fitted $Q_0 = 333 \mu\text{g}/\text{kg}$, $\alpha = 0.00548$, $R^2 = 0.881$; $P_{\max} = 52.94$, just beyond the tested price range). Total fentanyl intake did not differ significantly between male ($n = 7$) and female ($n = 7$) rats (Welch’s $t(7.0) = 0.23$, $p = 0.83$). Additionally, we did not observe significant changes in total locomotion across the four imaging sessions (one-way repeated-measures ANOVA, $F(3, 15) = 0.33$, $p = 0.80$, $\eta_G^2 = 0.026$), indicating the cross-session changes in demand were not driven by gross locomotor differences.

To ask whether these demand differences were reflected in DMS activity, we examined rewarded lever press-aligned population responses across the concentration series (Fig. 3.6d). The population response was largest and most sharply defined at the trained concentration ($70 \mu\text{g}/\text{mL}$), where the Σ Trace peaked at $2.13z$ roughly 1.1 s after the press, exceeding the response at any other concentration (peak $\leq 0.66z$). At the $222 \mu\text{g}/\text{mL}$ concentration, the press transient gave way to a brief suppression (trough $-0.85z$ near 2 s) and then a slow, delayed re-engagement that built after delivery offset and peaked several seconds later ($0.61z$ at ~ 6.8 s).

The response was small and tightly press-locked at $22 \mu\text{g}/\text{mL}$ ($0.66z$) and effectively absent under extinction ($0 \mu\text{g}/\text{mL}$; $0.30z$), where the heatmap dissolved into unstructured activity. Pooling neurons across all six animals confirmed that this trained-dose enhancement was not idiosyncratic to the representative animal: the population peak response was more than threefold larger at $70 \mu\text{g}/\text{mL}$ than at any other concentration ($0.67 \pm 0.06z$ vs. 0.14 ± 0.04 at 222 , 0.18 ± 0.05 at 22 , and 0.07 ± 0.05 at $0 \mu\text{g}/\text{mL}$; mean $\pm$ SEM, $n = 575\text{--}777$ neurons per concentration; Kruskal–Wallis $H = 49.2$, $p = 1.2 \times 10^{-10}$; trained vs. each, Mann–Whitney U , all $p \leq 8.9 \times 10^{-8}$; Fig. 3.6e).

Within each concentration, single-neuron press-aligned responses separated into two types: 1) A press-excited population and 2) A non-responsive population (Fig. 3.6f, g). Applying a single fixed responsiveness criterion across all concentrations, the fraction of press-responsive neurons did not differ significantly across the demand curve (Fig. 3.6h; $0.41\text{--}0.67$ across doses; Friedman $\chi^2(3) = 3.00$, $p = 0.39$), indicating that the size of the responsive population is conserved at the group level even as price varies.

Tracking the same neurons across the demand curve revealed that the conserved responsive fraction (Fig. 3.6h) masked a labile underlying membership. Gains and losses between groups were balanced so the fraction held near one-half even as roughly half of tracked neurons switched between press-responsive and non-responsive at each consecutive transition (55%, 57%, and 52% for $222 \rightarrow 70$, $70 \rightarrow 22$, and $22 \rightarrow 0$; Fig. 3.6h, i). This turnover was of equal magnitude when concentration was held constant across consecutive $70 \mu\text{g}/\text{mL}$ sessions, indicating a continuous reconfiguration of which neurons carry the press response rather than a fixed set.

To establish that this single-neuron turnover reflects genuine reorganization rather than tracking error, we compared cross-session agreement of the responsive class against the registration ceiling

the same pipeline achieves on ground-truthed data (Fig. 3.6j). Agreement between adjacent sessions was at chance (adjusted Rand index = +0.006; retention = 0.48 versus a 0.51 marginal responsive rate), far below the ≈ 0.94 overlap expected if responsive identity were stable. Thus, we conclude that the conserved press-responsive population (Fig. 3.6h) is continually rebuilt from a shifting set of neurons rather than a fixed ensemble. The turnover was of equal magnitude whether concentration rose, fell, or was held constant across consecutive 70 $\mu\text{g/mL}$ sessions (constant-dose ARI = +0.001), indicating ongoing session-to-session reorganization rather than a dose-locked code.

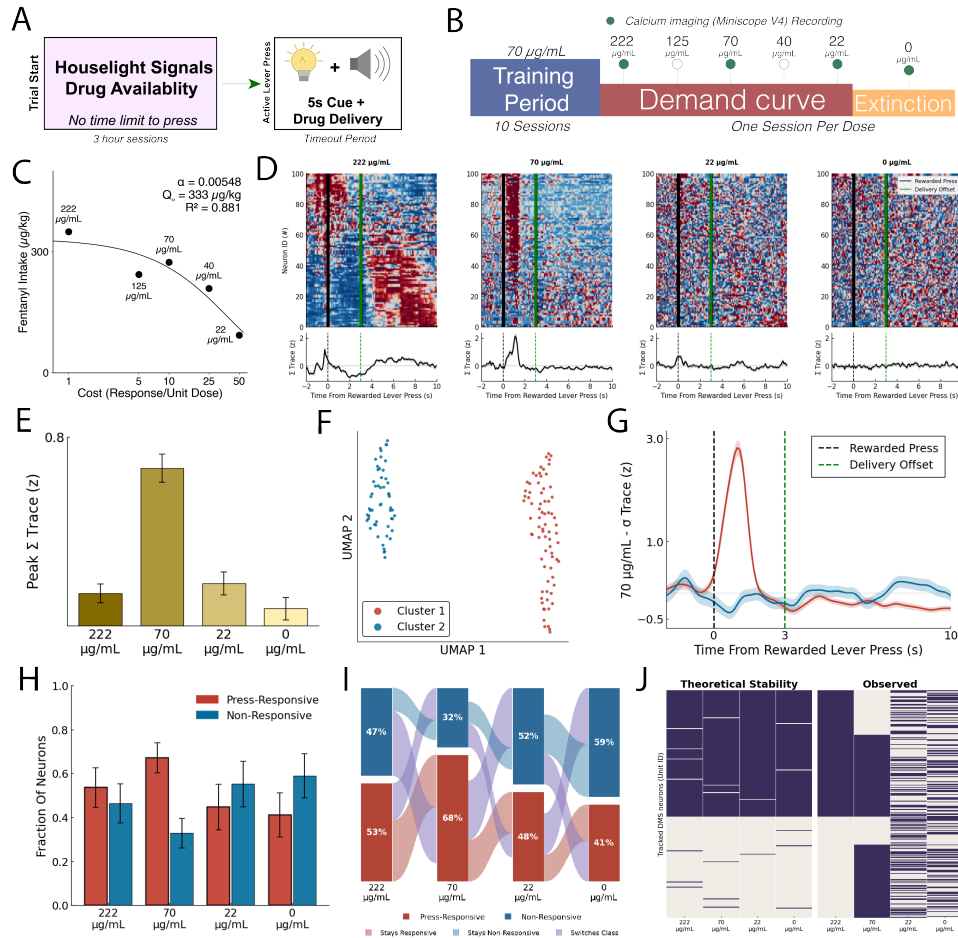


Figure 3.6: A conserved DMS population response masks near-complete single-neuron turnover across a fentanyl demand curve. **a**, Experimental timeline: habituation, acquisition at fixed concentration (70 $\mu\text{g/mL}$), and across-session demand-curve schedule with descending fentanyl concentrations: 222.0, 125.0, 70.0, 40.0, 22.0 $\mu\text{g/mL}$; volume held at 0.04 mL/delivery. **b**, Operant task schematic: oral fentanyl self-administration with active and inactive levers; deliveries scaled by per-session concentration to produce the unit-price gradient. **c**, Group demand curves plotted as consumption versus unit price on log-log axes; exponential demand model fit overlaid. **d**, Press-aligned DMS population activity for a representative animal at four fentanyl concentrations (222, 70, 22, 0 $\mu\text{g/mL}$). For each concentration, the heatmap shows the 100 most active neurons (per-neuron z -score, rows ordered by temporal center of mass; early responders at top) above the population summary trace (Σ Trace; mean $\pm$ SEM across all segmented neurons that session); heatmaps and traces are placed on a common scale across concentrations, and dashed lines mark rewarded-press onset ($t = 0$) and delivery offset ($t = 3$ s). **e**, Group quantification of the press-aligned response across all six animals via peak Σ Trace amplitude, shaded dark-to-light by descending concentration. **f**, UMAP embedding of single-neuron press-aligned responses for the representative animal at the trained concentration (70 $\mu\text{g/mL}$), each point a neuron colored by its Ward cluster. **g**, Cluster-mean response traces (mean $\pm$ SEM, per-neuron z) for the two clusters in **f**. **h**, Fraction of segmented neurons classed press-responsive versus non-responsive at each concentration. **i**, Session-to-session flow of response-class membership for S2C-tracked neurons across the demand curve (all animals pooled). Each column is a concentration, normalized to unit height; the two bands are press-responsive and non-responsive, labeled with their fraction of tracked cells. Ribbons connect cells tracked across each consecutive pair (222 $\rightarrow$ 70, 70 $\rightarrow$ 22, 22 $\rightarrow$ 0), colored by transition type: stays responsive, stays non-responsive, or switches class. **j**, Responsive-class identity across sessions for neurons S2C-tracked through all four imaging days. *Left* ("Theoretical Stability"): the raster the same tracking would produce if responsive identity were fixed, simulated by holding each cell's first-session label constant and flipping it only at S2C's per-session registration error rate ($\sim 1.5\%$). *Right* ("Observed"): the measured labels.

3.3 Discussion

We have presented Stars2Cells, a neuron tracking pipeline that adapts astrometric plate-solving to cross-session neuron registration in calcium imaging. By operating exclusively on centroid coordinates and using geometric descriptors invariant to rotation, translation, and uniform scaling, S2C sidesteps the requirements for spatial footprint data, activity traces, or manual alignment that limit existing methods.

The combination of accurate cross-session tracking and a GUI-driven distribution is what shifts longitudinal calcium-imaging analysis from a specialty technique into a standard option. Concretely, every chronic recording in an existing lab archive becomes a candidate for trajectory-level analysis: how do individual neurons reorganize their tuning across days of learning; do response profiles in reward-related circuits drift while the population code remains stable (Driscoll et al., 2017; Rule et al., 2019; Schoonover et al., 2021; Tiroshi et al., 2024); how does ensemble identity reorganize across the time-course of a neurodegenerative phenotype; what fraction of neurons recruited on day 1 of a behavioral paradigm are still engaged on day 30. Even though none of these questions are new, for most labs, they have been mechanically out of reach. Moreover, the underlying problem of point-cloud correspondence under rigid transform with partial overlap recurs in spatial transcriptomics, developmental biology, and wildlife re-identification, and S2C is released as a general tool of which neuron tracking is the first deployed application.

A priori, real-data validation of cross-session neuron tracking has a hard ceiling because a cell-resolved external ground truth cannot exist. For instance, optogenetic tagging of a labeled subpopulation establishes positive identity for the tagged cells but provides no false-positive characterization on the unlabeled majority that comprises the bulk of any tracker’s output. Dual-channel structural markers (e.g., tdTomato co-expressed with GCaMP) supply an identity signal across sessions but are partly self-referential when the structural channel also drives the registration step (Vergara et al., 2025a). Expert hand-annotation cannot integrate four-point geometric relationships across hundreds of nearly-identical neurons under arbitrary rotations and translations. The consequence is that real-data benchmarks for cross-session matching are convergent-evidence arguments, not direct measurements of matching precision and recall. They show that a tool produces consistent outputs where part of the answer is observable, but then they cannot score the answer the tool actually delivers. The synthetic benchmark with construction-time ground truth is the only scenario in which matching precision and recall are directly measurable, because the identity of every neuron is known by construction and every match the tool returns can be scored as correct or incorrect.

Therefore, we adopt this synthetic-benchmark approach as the load-bearing validation step, with parameter ranges informed by direct empirical observation of inter-session perturbations in our own and in published chronic miniscope recordings, and extended to bracket the more challenging conditions current hardware and surgical practice can produce. Astrometric plate-solving in astronomy is validated identically: Astropy and Astrometry.net (Hogg et al., 2008; Lang et al., 2010) are scored against synthetically rotated star fields and ground-truthed catalogs, not against expert-by-eye matching. Thus, considering the 1,262-paired-run benchmark establishes that

S2C achieves $F1 = 98.4\%$; the DMS demand-curve application (Section 3.2.4) demonstrates that the pipeline carries through end-to-end on real chronic data. As a further cross-modality check, we applied S2C unmodified to 431 Allen Brain Observatory two-photon containers, each a single cortical field of view imaged across three sessions. S2C reproduced the resource’s own cross-session registration (median $F1 = 0.97$; Section 3.4.8), providing convergent evidence rather than a direct measurement of matching accuracy.

Activity similarity cannot serve as that identity signal either. Matching neurons across sessions by activity similarity assumes away the very signal longitudinal experiments exist to measure. If cells are paired because their activity is alike, one can no longer ask whether that activity changed. Drift, plasticity, learning, and addiction paradigms are designed precisely to ask how a given neuron’s activity evolves across days, and thus, geometry and not activity must carry identity (Section 3.2.2, Fig. 3.3). Among existing tools, CellReg is the only like-for-like comparator given it alone takes the same input as S2C (per-session centroids/ROIs) and targets the same task (cross-session correspondence on one-photon miniscope data). The remaining tools either build on CellReg’s spatial-correlation backbone or take different inputs to answer a different question (Fig. 3.1c–d, Supplementary Note 7). Head-to-head, S2C outperformed CellReg across every benchmark condition. Furthermore, this gap is not an artifact of degraded footprints given that even when granted perfectly stable, best-case footprints, ROI matching still could not recover from the inter-session FOV variability chronic single-photon recordings routinely produce (Fig. 3.4g). Having validated S2C on synthetic ground truth and shown it to outperform its only like-for-like comparator for reasons structural rather than fixable, we now apply it to real recordings.

Applying S2C to DMS imaging during oral fentanyl self-administration surfaced biologically meaningful structure that would have been inaccessible to prior tools, retaining roughly three times as many longitudinally trackable neurons as ROI matching (Section 3.2.4, Fig. 3.5f). That retention exposed a stable group-level press response whose single-cell membership turns over completely between sessions (Section 3.2.5, Fig. 3.6h–j). This is the signature of representational drift established in hippocampus, cortex, and piriform, now be extended to the striatum (Ziv et al., 2013a; Driscoll et al., 2017; Schoonover et al., 2021; Rule et al., 2019; Domingues et al., 2025). Moreover, geometry-carried identity is what made the distinction measurable given that activity-based matching would assume the identity it exists to test, and ROI matching misses the pairs outright. Thus, since a bulk signal sums across cells, this turnover is invisible at the aggregate level and requires single-cell imaging with same-cell tracking. Pairing this tracking with D1/D2 reporters would resolve whether the cells cycling through the responsive group do so within or across striatal pathways, a natural next step beyond this proof-of-concept.

Several limitations bound S2C’s scope. The software’s accuracy depends on centroid extraction quality from the upstream pipeline; systematic localization errors will propagate. The current implementation assumes rigid (or similarity) inter-session transforms; large nonlinear distortions are handled only to the extent they appear locally rigid within quad neighborhoods. The descriptor pipeline is $O(NkK^2)$ (linear in N at fixed k and K) and descriptor matching is $O(N^2)$, the latter

the binding cost at high density. We tested 100–1,000 neurons per FOV (the standard range for chronic miniscope and two-photon cortical imaging), but mesoscope-class recordings of 3,000–12,000+ neurons (Stringer et al., 2019; Sofroniew et al., 2016; Stirman et al., 2016) are outside this scope by design, where the 4D descriptor space saturates and false-positive matching rises. Adapting S2C to that regime (locality-sensitive hashing in descriptor space, hierarchical RANSAC, GPU batching of inlier evaluation) is a tractable extension rather than a new method, which we have not pursued here because the dominant adoption gap sits at densities S2C already covers. Regardless, within the tested ranges, S2C did not drop below 93.6% F1 on any condition, and we therefore cannot point to an empirically observed sub-90% breaking point. Two failure modes are predictable from the algorithm’s construction, however, and worth flagging for users whose preparations sit outside the chronic-rodent-imaging scope this paper addresses. First, when nearest-neighbor distance approaches the centroid jitter floor, then quads from different neighborhoods become descriptor-similar and the matching step’s signal-to-noise floor could break down. Second, because our RANSAC step fits rigid transforms (rotation + translation) with scale held fixed, it rejects genuine matches in preparations where tissue expands uniformly between sessions, such as developmental two-photon recordings, organoids, and plant-growth tracking. However, releasing the scale parameter inside RANSAC is a single-line code change, but we have not implemented or validated it since it is outside the scope of the present paper. Future work could also incorporate locally affine models or deep-learned descriptors to handle nonlinear distortions more explicitly; the higher-density extension is outlined above.

Thousands of investigators now image neurons in behaving animals, but only a small minority can follow the same cells across days. Stars2Cells adds that capability on top of the segmentation pipelines these recordings already rely on. Thus, once individual neurons are linked across sessions, a chronic recording stops being a set of isolated snapshots and becomes a record of how the same cells change over an untold number of modalities.

3.4 Methods

3.4.1 Stars2Cells algorithm

S2C solves the same problem astronomers solve when plate-solving a photograph of an unknown patch of sky: given a sparse set of points in unknown rotation and translation, with some points missing and some new ones appearing, identify which point is which. The solution is borrowed from astrometric quad matching (Lang et al., 2010) and lent to neurons in the name S2C in that no individual point carries enough information to be identified, but four points carry much more. The relative geometry of any four neurons is invariant under rotation and translation, and once that geometry is encoded as a single vector in $\mathbb{R}^4$, two sessions’ quads can be compared directly without ever solving for the unknown transform. Once enough quads agree, the transform falls out of them.

S2C was implemented in Python 3.10 against a pinned dependency stack: NumPy 1.26.4 and SciPy 1.13.0 (numerical core, installed via Conda), faiss-cpu (cosine-distance descriptor matching), scikit-learn 1.8, scikit-image 0.26, NetworkX 3.6, tqdm 4.67, psutil 7.0, and PyQt5 5.15 / PyQt5-sip 12.18 / pyqtgraph 0.14 for the GUI; image and array I/O via Pillow 12.1, ImageIO 2.37, tifffile 2026.3.3, and Matplotlib 3.10. The full pinned requirements.txt ships with the source release. Both builds are produced with PyInstaller, which bundles a self-contained Python runtime and all scientific dependencies into the application itself: the macOS build is delivered as a signed and notarized DMG containing Stars2Cells.app, and the Windows build is delivered as a ZIP archive containing a standalone Stars2Cells.exe. End users do not need to install Python, manage Conda or virtual environments, edit PATH, or use the command line at any point in the workflow. Standard operating-system gatekeeping warnings (macOS Gatekeeper, Windows SmartScreen) may appear on first launch and are dismissed through the OS’s standard install-confirmation flow, requiring no command-line intervention.

Step 1: Quad descriptor generation. Quads are not enumerated brute-force – $\binom{N}{4} \approx 2.6 \times 10^9$ candidates at $N = 500$ is intractable, and almost all of them are uninformative (very thin or very small quads carry essentially no positional information). S2C builds quads diagonal-first and draws each quad exactly once.

The diagonal set per session is constructed from $k_{\text{local}} = \min(15, N - 1)$ (tunable) KD-tree nearest neighbors plus $k_{\text{random}} = \max(2, k_{\text{local}}/2)$ random long-range partners (tunable) per neuron (with KNN exclusions, seeded for reproducibility). Local diagonals stabilize the descriptor under small inter-session shifts; long-range diagonals produce FOV-spanning quads with small descriptor blur ($\sigma_d \propto 1/L_q$, where L_q is the longest pairwise distance of the quad) and dominate matching when the imaging field shifts substantially between sessions. For each unique diagonal (i, j) with $i < j$, the signed perpendicular height of every other neuron k is computed inline via the 2D cross product:

$$h_k = \frac{(\mathbf{P}_j - \mathbf{P}_i) \times (\mathbf{P}_k - \mathbf{P}_i)}{\|\mathbf{P}_j - \mathbf{P}_i\|} \quad (3.1)$$

where $\mathbf{P}_i \in \mathbb{R}^2$ is the centroid coordinate of neuron i and $\times$ denotes the scalar 2D cross product

($\mathbf{u} \times \mathbf{v} = u_x v_y - u_y v_x$). The top- K third points by $|h_k|$ are retained ($K = 25$ default) and combined into $\binom{K}{2}$ candidate quads per diagonal. For each quad $\{i, j, C, D\}$, the four-dimensional descriptor is constructed by projecting C and D onto the diagonal-aligned coordinate frame:

$$\mathbf{q} = \left(\frac{\mathbf{AC} \cdot \hat{\mathbf{u}}_x}{L}, \frac{\mathbf{AC} \cdot \hat{\mathbf{u}}_y}{L}, \frac{\mathbf{AD} \cdot \hat{\mathbf{u}}_x}{L}, \frac{\mathbf{AD} \cdot \hat{\mathbf{u}}_y}{L} \right) \quad (3.2)$$

where $\mathbf{A}, \mathbf{B}$ are the two endpoints of the quad’s longest pairwise edge, $L = \|\mathbf{B} - \mathbf{A}\|$, $\hat{\mathbf{u}}_x = (\mathbf{B} - \mathbf{A})/L$, and $\hat{\mathbf{u}}_y$ is its 90° counter-clockwise rotation. Canonical ordering enforces $x_C \leq x_D$ (swap labels otherwise) and $(x_C + x_D) \leq 1$ (axis flip otherwise); these two rules are what convert the raw projection from rotation-equivariant to fully rotation-invariant. Degenerate quads with $|y_C|, |y_D| < 10^{-4}$ (collinear) are discarded.

The same quad has six pairwise edges and could in principle be reached from any of them. The **ownership guard** emits a quad from a diagonal only when that diagonal equals the quad’s longest edge, evaluated within a 10^{-6} relative tolerance. This yields exactly one emission per quad with no global deduplication pass. For real-valued centroid coordinates, two distinct pairwise distances are equal only on a measure-zero set, so each quad has a unique longest edge almost surely; the 10^{-6} relative tolerance in the ownership test collapses the rare floating-point near-tie to a single emission.

Coverage remediation. The ownership guard can miss a neuron in a dense region where every quad incident to it gets claimed by some longer neighboring diagonal, and thus, the neuron contributes zero descriptors. To solve this, after the main pass, S2C identifies neurons whose quad count falls below a tunable parameter (default = 0.4) multiplied by the median(coverage) over the field and re-runs descriptor generation on their incident diagonals with the ownership guard disabled. The remediation batch is deduplicated against the existing quad set on sorted vertex tuples and quality-pruned by $|y_C| + |y_D|$ (keeping the top fraction, default 1.0 – tunable). Formal detection criteria and the uniqueness handling for the remediation batch are given in Supplementary Note 3.

Descriptor blur and saturation. Per-neuron centroid jitter of standard deviation σ_j pixels propagates to descriptor space as $\sigma_d \approx \sqrt{2}\sigma_j/L_q$, where L_q is the quad’s diagonal length: each projected descriptor coordinate involves a difference of two jittered points (the $\sqrt{2}$ factor) normalized by L_q . Longer-range diagonals therefore produce smaller descriptor blur, which is why the diagonal set mixes KNN-local edges with random long-range partners. The descriptor field is *saturated* when median nearest-neighbor distance in descriptor space falls below median descriptor blur, $\text{median}(\text{NN}_d) < \text{median}(\sigma_d)$; in that scenario jitter-induced descriptor noise exceeds inter-descriptor separation and false matches are unavoidable regardless of threshold. S2C evaluates this criterion at the end of Step 1 and emits a warning if saturation is detected.

Step 1.5: Threshold calibration. Whether two quads count as the same comes down to a distance threshold in $\mathbb{R}^4$ descriptor space: tight enough to suppress chance collisions between unrelated quads, but loose enough to absorb the centroid jitter that survives MPS preprocessing. Both moving parts depend on N : as the field densifies, the local-diagonal lengths that anchor same-cell descriptors shrink, so the same centroid jitter occupies a larger fraction of each diagonal

and same-cell descriptor blur grows. The threshold that must clear that blur therefore grows with field size.

For each animal, S2C samples a fixed number of session pairs (default 10k quads per pair – tunable) and sweeps similarity thresholds $\tau \in [0, 1]$ using descriptor distance only. At each τ , raw descriptor matches are passed through a consistency filter, and a quality score is computed from the raw and filtered counts. The optimum τ^* is selected as the location of the per-pair quality peak. The relationship

$$\tau(N) = C \cdot \sqrt{N} \quad (3.3)$$

is then fit by least squares across all sampled pairs from that animal, yielding C , C_{std} , and R^2 . The $\sqrt{N}$ form follows from a geometric argument (Supplementary Note 2): local-diagonal length scales as $L_q \propto N^{-1/2}$, so same-cell descriptor blur scales as $\sqrt{N}$, and τ is bounded below by it. Because the derivation assumes uniform density, the regression R^2 reports how closely a field approximates uniform packing, not calibration quality: a lower R^2 simply means centroids are non-uniformly spaced (clustering or density gradients, both expected), with the residual imprecision absorbed downstream by the consistency filter and RANSAC.

Step 2: Descriptor matching. For each session pair (R, T) , every quad descriptor in T is matched to its top-1 nearest neighbor in R under cosine distance via FAISS (IndexFlatIP with L2-normalized descriptors), and matches with distance below the per-animal calibrated τ are retained. A geometric consistency filter then rejects matches whose ratio of inter-quad centroid distances is inconsistent with the bulk distribution (consistency threshold = 0.8 – tunable). Surviving matches are written to a lightweight NPZ file containing the matched quad indices and the underlying centroid coordinates.

Step 2.5: RANSAC geometric verification. The output of Step 2 contains real correspondences and chance-collision noise in unknown ratio. The structure that distinguishes them: every real match is consistent with a single rigid transformation $(\mathbf{A}, \mathbf{t})$ between the two sessions, and every fake match is not. S2C draws $K = 1,000$ (tunable) random subsets of $m = 4$ matched centroid pairs each, and estimates all K rigid transforms simultaneously via batched SVD on the cross-covariance matrices:

$$\mathbf{H}_k = \tilde{\mathbf{S}}_k^\top \tilde{\mathbf{D}}_k, \quad \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top = \text{SVD}(\mathbf{H}_k), \quad \mathbf{R}_k = \mathbf{V}_k \mathbf{U}_k^\top \quad (3.4)$$

where $\tilde{\mathbf{S}}_k$ and $\tilde{\mathbf{D}}_k$ are the centered source and destination samples of the k -th subset. Reflections are corrected by flipping the last column of $\mathbf{V}_k$ whenever $\det \mathbf{R}_k < 0$, after which the rotation is recomputed. Translation is recovered as $\mathbf{t}_k = \bar{\mathbf{D}}_k - \mathbf{A}_k \bar{\mathbf{S}}_k$ where the bars denote sample means. Inlier counts for all K candidate transforms are evaluated against all N matched pairs in float32 chunks sized to ~ 200 MB of working memory, with squared-residual threshold $(5.0 \text{ px} - \text{tunable})^2$. The transform with the most inliers is then re-estimated on its full inlier set, and matches with refined residual $> 5.0 \text{ px}$ (tunable) are rejected. The RANSAC residual threshold used is persisted into the Step 2.5 NPZ output, where it is consumed by Step 3 to set its distance cutoffs.

Step 3: Hungarian assignment and consolidation. Step 2.5 leaves a set of *quad*-level

correspondences that are geometrically consistent with the inferred rigid transform. Step 3 turns these into a one-to-one *neuron* mapping. Each surviving inlier quad votes for all $4 \times 4 = 16$ implied neuron pairings, building a vote matrix $\mathbf{V} \in \mathbb{N}^{N_R \times N_T}$. Raw vote counts are degree-normalized to remove a hub-neuron bias:

$$\tilde{V}_{ij} = \frac{V_{ij}}{\sqrt{d_i^{(R)} \cdot d_j^{(T)}}} \quad (3.5)$$

where $d_i^{(R)}$ and $d_j^{(T)}$ are the total quad degrees of neurons i and j in their respective sessions. The cost matrix combines vote-cost and spatial residual under the inferred transform:

$$C_{ij} = (\tilde{V}_{\max} - \tilde{V}_{ij}) + \lambda \cdot \|(\mathbf{A}\mathbf{p}_i^{(R)} + \mathbf{t}) - \mathbf{p}_j^{(T)}\| \quad (3.6)$$

with λ chosen so that the median spatial term equals the median vote term over voted entries ($\lambda = \tilde{C}_{\text{vote}}/\tilde{D}_{\text{spatial}}$), and $C_{ij} = \infty$ outside a distance cutoff of $3.0\times$ (tunable) the RANSAC residual threshold. Pairs that are spatially nearby but unvoted are not blocked outright; they receive a distance-only fallback cost above the maximum voted cost, so they remain assignable when no voted competitor exists.

The cost matrix is then padded with rectangular dummy rows and columns (one dummy slot per real neuron in each direction, each priced at the global median voted cost) and solved with SciPy’s linear sum assignment routine (Pass 1). Without the dummies the Hungarian algorithm would return a complete matching that pairs every neuron to some partner; the fixed-cost dummy slots instead give each neuron an opt-out, so it accepts a real partner only when that partner is cheaper than the median dummy cost and is otherwise left unmatched, which is the correct behavior for dropouts and newly appearing cells that have no true correspondent. Pass-1 matches with post-hoc spatial residual exceeding $1.0\times$ the RANSAC threshold are rejected. Pass 2 then re-solves a smaller Hungarian over the unmatched residual ($N_R^{\text{unmatched}} \times N_T^{\text{unmatched}}$ submatrix) under a relaxed cutoff of $2.0\times$ the RANSAC threshold, with submatrix dummy cost set to the 75th percentile of finite submatrix costs (tunable). This recovers correspondences that Pass 1 conservatively rejected.

Each surviving match (i, j) is assigned a confidence score in $[0, 1]$ that averages four equally-weighted components: the degree-normalized vote score $\tilde{V}_{ij}/\tilde{V}_{\max}$; spatial proximity $\max(0, 1 - d_{ij}/d_{\text{cutoff}})$; assignment margin $(C_i^{(2)} - C_i^{(1)})/C_i^{(2)}$ where $C_i^{(1)}$ and $C_i^{(2)}$ are the best and second-best finite costs in row i ; and pass score (1.0 for Pass 1, 0.5 for Pass 2). Pairwise neuron correspondences are then chained transitively into multi-session tracks: a track threading neurons $n_1 \rightarrow n_2 \rightarrow \dots \rightarrow n_S$ across S sessions inherits the minimum link confidence as its conservative track confidence, and the mean as its expected confidence.

3.4.2 Temporal correlation control experiment

To test whether calcium-trace correlation could serve as a stand-alone identity signal, we generated 1,000 simulated neurons placed at identical spatial positions (Poisson-disk, $r_{\min}$ as in the main benchmark) across 5 sessions. For each session, calcium traces were independently simulated using

an AR(1) model: per-neuron Poisson spike trains with rates $\sim \mathcal{U}(0.05, 0.5)$ Hz at 10 Hz frame rate were convolved with a single-exponential calcium kernel ($\tau = 1$ s) using a discrete IIR filter with coefficient $a = e^{-1/(\tau \cdot \text{fps})}$, scaled by per-neuron amplitudes $\sim \mathcal{U}(0.5, 3.0)$, and corrupted with Gaussian noise ($\sigma = 0.05$). The random seed for trace generation differed between sessions, so the dynamics of any given neuron were statistically independent across sessions even though its position was perfectly preserved. Recording durations of 10, 20, 60, and 180 min were tested (6,000 / 12,000 / 36,000 / 108,000 frames). For each pair of sessions, traces were z -scored row-wise (in float32) and the full $1,000 \times 1,000$ Pearson correlation matrix was computed in row-chunks of 100 to avoid the ~ 8 GB peak of a single dense matmul; each neuron in session s_1 was assigned to its maximally correlated counterpart in session s_2 . Accuracy was scored against the (preserved) ground-truth identity. Across all 10 session pairs and all four durations, mean accuracy was indistinguishable from chance ($1/N = 0.1\%$).

3.4.3 Synthetic benchmark generation

Benchmark data were generated by a custom Python script that placed neurons via Bridson Poisson-disk sampling and applied deterministic affine transforms with NumPy under controlled per-condition seeds. For each combination of *animal* (8), *condition* (1–9 depending on tier; see below), *seed* (5; except Tier 0, single seed), and *neuron count* (100, 250, 500, 1,000), five consecutive sessions were generated from a base centroid field placed in a 600×600 -pixel FOV with $r_{\min}$ set to 55% of the theoretical hex-packing maximum.

Per-session transformation pipeline. Each session is generated from the base field by a fixed transform sequence detailed in Supplementary Note 5 (cumulative rotation, translation, clip-to-FOV, dropout, per-neuron jitter, ID permutation), with ground-truth identity captured before permutation.

Neuron-count tier coverage. All four neuron counts (100, 250, 500, 1,000) run all four tiers (Tier 0 + A + B + C; 8 productive conditions plus 1 identity sanity-floor).

3.4.4 Spatial footprint synthesis for the ROI-based matching comparison

ROI-based matching requires pixel-level spatial footprints (ROI masks), not just centroids. To enable a like-for-like comparison without a lossy round-trip through real movies, we synthesized footprints directly from the same centroid fields used to evaluate S2C. Each neuron was rendered as a 2D elliptical Gaussian with semi-major axis σ_{major} , semi-minor axis $\sigma_{\text{minor}} = \sigma_{\text{major}} \cdot \epsilon$ (eccentricity $\epsilon \sim \mathcal{U}(0.6, 1.0)$), orientation $\theta \sim \mathcal{U}(0, \pi)$, and peak intensity $\sim \mathcal{U}(0.6, 1.0)$. The mean radius was scaled with neuron count to mimic empirically observed CNMF/CNMF-E footprint compression in dense fields: $\bar{\sigma} = 12$ px at $N = 100$, 10 px at 250, 8 px at 500, and 6 px at 1,000 – values consistent with the ~ 8 –15 px effective radii reported for one-photon miniscope CNMF-E extraction (Pnevmatikakis et al., 2016b; Zhou et al., 2018a) and the FOV-density scaling used by CNMF-E itself. Per-neuron radius variability was $\text{CV} = 0.15$, clipped to $[0.5\bar{\sigma}, 1.8\bar{\sigma}]$. Footprints were rendered with a 2.5σ crop radius for efficiency and stored both as compact per-neuron patches (with bounding-box

offsets) and as dense $N \times H \times W$ arrays consumable directly by CellReg in either Python (.npy) or MATLAB (.mat, with $H \times W \times N$ axis order) formats. Each footprint was L2-normalized.

Two scenarios. Footprints were generated under two scenarios designed to bracket ROI matching’s plausible operating range:

- **Best-case (no shape perturbation).** The footprint shape parameters of each neuron are fixed once per animal, are unique per neuron, and reused unchanged across every session, condition, and seed; only centroid *position* changes between sessions according to the benchmark’s geometric perturbations. This gives ROI matching the strongest possible spatial-correlation signal: a neuron’s footprint shape is a perfect identity tag.
- **Worst-case (complete shape perturbation).** Per-session, per-neuron Gaussian perturbations are added on top of the base shape parameters: σ jitter $\sim \mathcal{N}(0, 0.5 \text{ px})$ (independent on major and minor axes), orientation jitter $\sim \mathcal{N}(0, 5^\circ)$, peak-intensity jitter $\sim \mathcal{N}(0, 5\%)$. These magnitudes were chosen to model focus drift, motion-correction residuals, CNMF re-estimation noise, and slow expression/photobleaching changes that real footprints undergo across days.

The same per-animal base-shape parameters are reused across all S2C-vs-ROI-matching comparisons within each scenario, so the spatial signal that the ROI-based matching sees is held constant within a scenario and varies only through the explicitly imposed per-session perturbations.

3.4.5 ROI-based matching reimplementations

To enable evaluation across all 32 productive (condition $\times$ neuron-count) cells on a single workstation, we reimplemented CellReg’s core registration pipeline in Python (NumPy, SciPy). The reimplementations faithfully port the pipeline component-for-component, in the order Sheintuch et al. specify it:

1. **Footprint normalization and FOV alignment.** Each footprint is L2-normalized to unit sum (matching MATLAB’s `normalize_spatial_footprints.m`), and sessions are zero-padded to a common canvas. A pre-matching alignment stage (`align_images`) estimates the rigid-body transform that maximizes cross-correlation between each non-reference session’s footprint projection and the reference, then applies that transform to both footprints and centroids. We retained CellReg’s full alignment step so that the comparison is not unfairly handicapped by inter-session shifts the tool is designed to absorb upstream of matching.
2. **All-to-all neighbor distributions.** For each cell pair across sessions within a maximal centroid distance d_{max} (default 20 μm at 1 $\mu\text{m}/\text{px}$, so 20 px), the spatial Pearson correlation ρ_{spatial} is computed over the union of non-zero pixels (corr2-equivalent), and the centroid distance d is recorded. The pooled (ρ, d) distributions across all session pairs are accumulated for model fitting.

3. **Two parallel mixture models on $(1 - \rho)$ and on d .** A spatial-correlation model fits a *lognormal* same-cell distribution and a *Beta* different-cell distribution to the $(1 - \rho_{\text{spatial}})$ histogram via 100 EM iterations, with sigmoid smoothing on the same-cell component (matching MATLAB CellReg). A centroid-distance model fits a lognormal same-cell + sigmoid-by-linear different-cell mixture to d via nonlinear least squares. Each model produces a Bayes-rule lookup $P(\text{same})$ as a function of its metric, and the intersection (where $P(\text{same}) = 0.5$) is reported.
4. **Best-model selection.** `choose_best_model` picks the model with the lower combined cost $\text{cost} = \text{FP_rate} + \text{FN_rate} + \text{MSE}$.
5. **Per-pair P_{same} assignment.** Each candidate neighbor pair is given its P_{same} via nearest-bin lookup in the chosen model.
6. **Initial registration – single greedy pass.** The registry is seeded with all cells from session 1; for sessions $2, \dots, S$, each cell is greedily matched to its highest-similarity registered neighbor above the threshold (or registered as a new cell otherwise). Slot conflicts are resolved by keeping the higher-scoring assignment and displacing the loser to a new row. *This is not a Hungarian solve*: it is a one-pass nearest-neighbor seed (`initial_registration_*`).
7. **Final registration – iterative cluster_cells (≤ 10 iterations).** The greedy seed is refined by the iterative `cluster_cells` routine. Each iteration performs three sub-steps: (A) per-cell reassignment within $1.7 \times d_{\text{max}}$ using the maximal-similarity criterion (split clusters whose best similarity falls below threshold; move or swap cells between clusters when a better-fitting cluster is found); (B) cluster merging across non-overlapping session sets when the maximal pairwise P_{same} exceeds threshold; (C) cleanup of empty rows. The loop terminates when fewer than 10 reassignments occur in an iteration or after 10 iterations.
8. **Cell-level scoring.** Each registered cluster receives positive (mean P_{same} over confirmed pairs), negative (mean $1 - P_{\text{same}}$ over absent-session pairs), and exclusive (penalized for spurious extra neighbors) scores via `compute_scores`.

Configuration parameters. Microns per pixel = 1.0; maximal pairing distance = 20 μm ; alignment = “Translations + Rotations”; P_{same} threshold = 0.5. These settings match standard CellReg defaults for miniscope data.

Search-radius choice. The 20 μm radius defines the neighborhood within which CellReg’s two mixtures are fitted: the same-cell mode (lognormal on centroid distance; lognormal on $1 - \rho_{\text{spatial}}$) sits at small values and the different-cell mode is fitted across the remainder of the radius, with the discriminative power of P_{same} deriving from the gap between them. Extending the radius to admit additional candidate matches in the highest-displacement conditions would flood the neighbor pool with different-cell pairs at every distance, compress the gap between the two modes, and produce low-precision matches under an identity model that has ceased to mean what CellReg’s

authors specified. We therefore retained the standard miniscope-default radius across the full benchmark. When per-step inter-session displacement exceeded that radius, the same-cell mode of CellReg’s centroid-distance mixture lost within-radius support and only chance pairings between unrelated cells remained ($TP = 0$, $F1$ undefined); these runs were excluded from the paired statistic rather than recovered by parameter widening, and they concentrate in the B-drift condition, whose 5–25 px/step translation range crosses the 20 μm horizon directly. Because these excluded runs are precisely the cases where ROI-based matching failed completely, the exclusion is conservative with respect to our central comparison: had we instead scored them as $F1 = 0$, CellReg’s pooled $F1$ would fall and the S2C advantage would widen. Dropping them therefore understates rather than inflates the reported gap.

To verify the reimplementations, we generated held-out 2-session test sets at all four benchmark neuron counts ($N = 100, 250, 500, 1,000$) using `generate_cellreg_test.py`, each with ~ 1 px per-neuron centroid jitter and 10% session-2 dropout, with permuted ROI IDs and a separately stored ground-truth identity table mapping each session’s permuted indices back to base identity. Identical synthetic spatial footprints were exported in both formats: `.npy` dictionaries for the Python implementation and `.mat` arrays in $H \times W \times N$ orientation for native MATLAB CellReg (v1.5.9, which incorporates the FOV-alignment fix introduced in v1.5.5). Both pipelines were run end-to-end with matched parameters (microns/pixel = 1.0, maximal distance = 20 μm , $P_{\text{same}} = 0.5$, “Translations + Rotations” alignment, “Spatial correlation” model selection at both intermediate and final steps), and the resulting cell-to-index maps were compared row-by-row against ground truth.

At every neuron count tested ($N = 100, 250, 500, 1,000$), the two implementations agreed on *every* matched pair: identical model selection (spatial correlation), identical match count, identical pair assignments, identical singleton (session-1-only) classification of the dropped neurons, and identical 100% precision and 100% recall against ground truth. The set difference between Python and MATLAB matched-pair sets was empty in both directions across all four neuron counts. With pair-level concordance established across the full neuron-count range of the benchmark, the parallelized Python reimplementations were used for the full benchmark. Because the implementation is deterministic and its code path does not branch on input difficulty, the pair-level concordance established on these benign test sets carries over to the harder benchmark conditions; we reimplemented in Python solely because running native MATLAB CellReg across all 32 (condition $\times$ neuron-count) cells at scale on a single workstation was infeasible, not to alter its behavior.

Performance under benign conditions. Pair-for-pair agreement with native MATLAB CellReg is established in Methods (“ROI-based matching reimplementations”). A one-to-one reproduction of their reported numbers is not attainable: Sheintuch et al. report false-positive/false-negative rates rather than $F1$, the only public CellReg data are unlabeled (precluding scored precision/recall), and their real-recording error rate cannot be matched by clean synthetic data without tuning density and noise to the target, which is circular. We therefore rely on the agreement with the released MATLAB tool as the rigorous check. As a qualitative consistency test, on well-aligned synthetic

data the reimplementation registers cells with $<5\%$ error, consistent with the $<5\%$ rates Sheintuch et al. report (Sheintuch et al., 2017b).

3.4.6 Evaluation metrics

Precision, recall, and F1 were computed from ground-truth identity labels. A match was scored as true positive if both neurons shared the same base identity (accounting for permutation). For multi-session chains, metrics were computed pairwise across all four consecutive transitions ($A \rightarrow B$, $B \rightarrow C$, $C \rightarrow D$, $D \rightarrow E$). We selected F1 as the headline metric since both error modes carry biological cost in this setting: a false positive corrupts a neuron’s apparent dynamics by binding it to traces from a different cell, while a false negative discards a real cross-session correspondence and reduces the longitudinal sample. F1 weights both equally, which we considered preferable to weighted variants (e.g., F_β) that would require choosing β on grounds we have no principled basis to set.

3.4.7 Bibliometric adoption analysis

To quantify the longitudinal-tracking adoption gap reported in Fig. 3.1 and the Introduction, we performed a programmatic bibliometric survey using the NCBI Entrez API (BioPython 1.83) for publication counts and the OpenAlex API for tool-citation counts, on a date locked to manuscript submission. For the PubMed queries, all matching PMIDs were retrieved via `esearch`, full record summaries were pulled via `esummary` in 500-ID batches, and unique LastAuthor fields were tallied. Three measures were taken: (i) “two-photon labs” (all years, `two-photon[Title/Abstract]`), yielding 8,484 unique last authors from 19,642 unique PMIDs (full uncapped retrieval); (ii) “longitudinal labs” (2021+, `longitudinal[Title/Abstract] AND any of calcium imaging; two-photon; miniscope; one-photon; tracking neurons; cell registration` in Title/Abstract); and (iii) tool-specific adoption, estimated as citations of each tool’s canonical reference publication via the OpenAlex API, using manually verified DOIs. We report citation counts because no single index captures software adoption completely: PubMed title/abstract queries miss tools cited only in Methods (PubMed indexes no full text); open-access full-text indices (Europe PMC) recover Methods/Results mentions but only for the open-access subset; and citation counts, while covering the full literature irrespective of access status, conflate citation with use and under-count tools distributed without a citable paper. Counts are therefore transparent order-of-magnitude estimates; tools published within approximately one year of access (CaliAli, Track2p) are correspondingly under-represented and reflect recency rather than limited utility. Active-laboratory population estimates for UCLA Miniscope (~ 500 labs) and Inscopix nVista/nVoke systems (~ 600 labs) are taken as *manufacturer/ community estimates as of April 2026*, sourced from the UCLA Miniscope BRAIN Initiative ToolMakers page (BRAIN Initiative, 2025) and reporting on the Bruker–Inscopix acquisition (Photonics Media, 2022), respectively.

3.4.8 Cross-modality validation on the Allen Brain Observatory

To test whether S2C generalizes beyond one-photon miniscope data and beyond recordings from our own laboratory, we applied it without modification to publicly available two-photon calcium-imaging data from the Allen Brain Observatory Visual Coding 2P dataset (Allen Institute for Brain Science, 2016; de Vries et al., 2020). This resource images the same field of view in mouse visual cortex across three sessions (A, B, C) per experiment container and provides, for each session, per-ROI segmentation masks together with a `cell_specimen_id` that links the same neuron across sessions via the Allen Institute’s own cross-session registration. We retrieved 431 experiment containers with a complete A/B/C session set through the `AllenSDK BrainObservatoryCache`, reduced each session’s ROI masks to centroids (mask center of mass), and ran the unmodified S2C pipeline on the centroids alone, withholding all `cell_specimen_id` labels so that registration was blind to Allen’s correspondence. For each container we scored S2C’s $A \rightarrow B$ and $B \rightarrow C$ assignments against the shared-`cell_specimen_id` pairs and report F1 (Evaluation metrics). A session pair was scored only when both sessions contained enough neurons to form quad descriptors, leaving $n = 423$ scorable $A \rightarrow B$ pairs and $n = 419$ $B \rightarrow C$ pairs. Moreover, given that the Allen registration is itself a spatial-footprint method rather than cell-resolved ground truth, we treat this agreement as convergent evidence of cross-modality transfer rather than a direct measurement of matching accuracy (Discussion). Across these containers S2C reproduced the Allen registration with median $F1 = 0.97$ ($A \rightarrow B$; IQR 0.46–1.00) and 0.97 ($B \rightarrow C$; IQR 0.47–1.00).

3.4.9 Animals and housing

For all experiments, 8-week-old male and female Long-Evans rats were bred in-house. Rats were dual-housed in a temperature- and humidity-controlled room on a 10:14 dark-light cycle and acclimated to the facility for ~ 1 week before beginning experimental procedures at 9–10 weeks old. Food and water were available ad libitum throughout the experiment (weight range: 250–450 g). Experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of the VA Puget Sound Health Care System and complied with the NIH Guide for the Care and Use of Laboratory Animals. The study is reported in accordance with the ARRIVE 2.0 guidelines.

ARRIVE 2.0 reporting elements. A total of $n = 6$ animals (3 male, 3 female) were used in the imaging cohort; this sample size was chosen to match the per-cohort n used in the precedent oral fentanyl behavioral-economics work in the same lab (Coffey et al., 2023) and is treated here as a proof-of-concept demonstration of the MPS $\rightarrow$ S2C stack on real chronic miniscope data rather than as a powered biological-discovery experiment. Behavioral sessions were run in fixed-time daily blocks; imaging-day order was fixed by the demand-curve schedule (sessions 12, 14, 16, and 17). *Exclusion criteria, pre-specified:* animals were excluded from analysis if (i) viral expression was outside the dorsomedial striatum on histology, (ii) the GRIN-lens placement was outside the dorsomedial striatum on histology, (iii) baseline operant responding failed to stabilize during acquisition, or (iv) the miniscope FOV was unusable (lens occlusion, severe motion artifact uncorrectable by MPS, or

signal loss).

3.4.10 Oral fentanyl self-administration and behavioral economics

Animals ($n = 7$ male, $n = 7$ female) were trained on an oral fentanyl self-administration paradigm following the procedure of (Coffey et al., 2023), with concentrations and volumes adjusted for the behavioral-economics demand-curve schedule used here; of these, a subset of 6 ($n = 3$ male, $n = 3$ female) were implanted with a GRIN lens and miniscope baseplate and constitute the imaging cohort analyzed throughout (see “Surgical preparation and chronic miniscope imaging”). Training was conducted at a fixed concentration of $70 \mu\text{g/mL}$, with each delivery providing 0.04 mL of solution. After acquisition and stable responding on the training concentration (sessions 5–9: 56.0 ± 9.2 active-lever presses per session, mean $\pm$ SEM, $n = 14$; Friedman $\chi^2(4) = 2.79$, $p = 0.59$), animals advanced to an across-session demand-curve schedule, in which the fentanyl concentration descended across days while the response cost per unit dose increased monotonically. The per-session concentrations (in $\mu\text{g/mL}$) were 222.0, 125.0, 70.0, 40.0, and 22.0, followed by extinction at $0 \mu\text{g/mL}$, with delivery volume held constant at 0.04 mL .

For each delivery, the dose in micrograms per kilogram was calculated as the per-session fentanyl concentration multiplied by the per-delivery volume (0.04 mL) and divided by the animal’s body weight in kilograms. Whole-brain fentanyl concentration over time was estimated using the same two-compartment model previously applied to oral fentanyl in rat (Coffey et al., 2023), adapted from the cocaine pharmacokinetics methodology of (Pan et al., 1991) with a larger elimination coefficient ($k_{\text{el}} = 0.175$) to account for fentanyl’s relatively longer half-life. Briefly, each delivery contributed a bi-exponential term with a fast distribution component and a slower elimination component, parameterized by a transfer constant k , an apparent brain volume v , and rate constants α and β . In practice, this was implemented as a convolution of the delivery times with a kernel proportional to $\exp(-\alpha\Delta t) - \exp(-\beta\Delta t)$, where Δt is the time since each prior delivery. Concentrations were evaluated on a dense temporal grid spanning the full 180-minute session for each animal and then averaged across animals to obtain a population-mean concentration trajectory for each demand-curve session.

3.4.11 Pose estimation and kinematic analysis

Behavioral videos were manually annotated using a custom-built Python graphical interface (Python 3.10, OpenCV 4.8, Tkinter) designed for pose annotation. A custom annotation interface was used to ensure precise control over keypoint definitions, consistent rat-centered cropping in low-light operant chambers, tight alignment with trial structure, and curated sampling of behaviorally relevant frames (e.g., stretch-attend, approach, withholding), producing cleaner training data and more reliable kinematic extraction than would be obtained with generalized pose-tracking frameworks. For each video frame, seven anatomical keypoints were labeled: head center, body center, tail, front right paw, front left paw, left back paw, and right back paw. The annotation tool provided frame-by-frame navigation, drag-and-drop point placement, interpolation between labeled frames, and quality control

features including frame-by-frame review and correction capabilities. Annotations were exported in YOLO pose format, with each labeled frame producing: (1) a cropped image containing the rat, and (2) a corresponding text file specifying normalized bounding box coordinates (x_center , y_center , width, height) and keypoint positions (x , y , visibility) for all seven anatomical landmarks. Pose estimation models were trained using the YOLOv8n-pose architecture (Ultralytics YOLOv8) initialized from pretrained COCO-pose weights. The training dataset was split 80/20 into training and validation sets. Models were trained for 5 epochs with early stopping (patience = 3 epochs) using the following hyperparameters: image size = 640x640 pixels, batch size = 16, learning rate = 0.001 (with cosine decay), augmentation enabled (rotation $\pm 10^\circ$, translation $\pm 10\%$, scale $\pm 10\%$). Model performance was evaluated using mean average precision at IoU thresholds of 0.5 (mAP50) and 0.5:0.9 (mAP50-95) for both bounding box and keypoint detection. Final model selection was based on validation set mAP50-95 for keypoints. The best-performing model achieved mAP50 = 0.988 and mAP50-95 = 0.914 on the held-out validation set. Trained models were applied to all behavioral videos to generate frame-by-frame predictions of the seven keypoints. For each detection, the model returned (x , y) coordinates and confidence scores (0-1) for each keypoint. Predictions were retained only if: (1) detection confidence exceeded 0.5 for the bounding box, and (2) all seven keypoint confidence scores exceeded 0.5.

Trial-level movement features were extracted from tracked keypoint trajectories aligned to behavioral event timestamps from the operant system. To test whether gross motor output changed across the imaged sessions, we computed a session-level locomotion measure independent of trial structure. For each session, total locomotion was defined as the cumulative path length of the body-center keypoint via the sum of frame-to-frame Euclidean displacements over all retained frames of the session (pixels; chamber and camera geometry were held constant across the compared sessions, so pixel units are directly comparable across them). Frames failing the confidence criteria above were linearly interpolated before summation so that detection gaps did not inflate displacement. Total locomotion was compared across sessions with a one-way repeated-measures ANOVA (within-subject factor: session; $n = 6$ rats).

3.4.12 Surgical preparation and chronic miniscope imaging

Fourteen days before behavior, rats ($n = 3$ male, $n = 3$ female) were anesthetized with 1–3% isoflurane and placed in a custom robotic stereotaxic instrument (Coffey et al., 2013). The skull was leveled to bregma-lambda within ± 0.05 mm. A small craniotomy was made above right dorsomedial striatum (DMS). Animals received a 1,000-nL injection of AAV1-EF1 α -jGCaMP8f-WPRE (Zhang et al., 2023a) (Addgene #176756) at 100 nL/min via a 33-gauge needle (NF33BV-2, WPI). Coordinates: AP -0.3 mm, ML +2.4 mm, DV -4.5 mm. After 10 min of diffusion, the needle was withdrawn. A GRIN lens (1 mm diameter, 9 mm length; GRINtech) was lowered to DV -4.4 mm and secured with jeweler’s screws, cyanoacrylate, and dental acrylic (Parkell C&B Metabond). Animals received meloxicam (2.5 mg/kg s.c.) post-operatively and were monitored for ≥ 3 days.

Calcium imaging was conducted using the Miniscope V4 (UCLA; www.miniscope.org) where,

prior to the demand-curve start, animals underwent two dedicated habituation sessions on consecutive days in which they were connected to the miniscope and allowed to freely explore the chamber, ensuring acclimation to the head-mounted setup and minimizing handling- or equipment-related novelty during imaging days. The miniscope was connected via lightweight coaxial cable to the data acquisition system (Miniscope DAQ v3.2); video streams were saved as .avi with synchronized behavioral timestamps. A custom system for cable protection was used, consisting of (1) strain-relief epoxy at the baseplate–cable interface and (2) a flexible wire cage surrounding the coaxial cable from headstage to commutator. This was essential for recordings during high-motion and aversive-stimulation sessions.

3.4.13 Miniscope Processing Suite (MPS) pipeline

All calcium imaging data were processed using the Miniscope Processing Suite (Peden-Asarch et al., 2026c), our open-source GUI-based pipeline optimized for long-duration one-photon imaging. Processing parameters were held consistent across animals, with automatic parameter suggestions used to optimize performance for individual datasets. The pipeline performs preprocessing (background subtraction, spatial filtering, rigid translational motion correction, automatic detection and removal of erroneous frames), interactive field-of-view cropping, NNDSVD-based dimensionality reduction, watershed-based cell detection and initialization, constrained non-negative matrix factorization (CNMF) with autoregressive calcium dynamics for temporal refinement, dilation-bounded localized LASSO regression for spatial refinement, and a final temporal convergence pass with quality filtering on size, signal-to-noise ratio, and spatial-temporal correlation. Processed datasets were exported as compressed npy files containing spatial footprints (A), temporal components (C), deconvolved spike estimates (S), and background terms (b, f) with full metadata. All spatial footprints and calcium traces were manually inspected using the MPS Data Explorer interface and excluded if they exhibited vascular morphology, footprints outside the field of view, non-stationary baseline dynamics inconsistent with calcium transients, or anatomical localization outside the dorsomedial striatum based on histological verification. Full pipeline parameters and component-by-component validation are reported in (Peden-Asarch et al., 2026c). Additionally, per-session yield varied across animals (one-way ANOVA, $F(5, 30) = 4.22$, $p = 0.005$; $\eta^2 = 0.41$), consistent with expected animal-to-animal differences in viral expression and imaging-field placement, but was stable across sessions within animal, showing no systematic drift over the imaging window (Friedman $\chi^2(5) = 6.42$, $p = 0.27$; Spearman $\rho = -0.12$, $p = 0.47$).

3.4.14 Histology and viral injection confirmation

Following the conclusion of behavioral testing, animals were deeply anesthetized and transcardially perfused with phosphate-buffered saline followed by 4% paraformaldehyde. Brains were extracted, post-fixed overnight, and cryoprotected in 30% sucrose. Coronal sections (40 μm) through the dorsomedial striatum were collected on a freezing microtome and immunostained for jRCaMP8f expression, with lens placement verified from the implant track. Sections were incubated with

primary antibody chicken anti-GFP (1:1000, Aves Labs Cat#GFP-1020, RRID:AB_10000240) and detected with secondary antibody Alexa Fluor 488–conjugated goat anti-chicken IgY (1:500, Thermo Fisher Scientific Cat#A-11039, RRID:AB_2534096). Sections were counterstained with DAPI, mounted, and imaged on a Keyence fluorescence microscope. Animals with viral expression or lens placement outside the dorsomedial striatum were excluded from analysis.

3.4.15 Demand-curve fitting and derived behavioral-economics parameters

Per-session fentanyl consumption was expressed as total intake ($\mu\text{g}/\text{kg}$), computed as the number of reinforced deliveries in a session multiplied by the per-delivery dose defined above. For each animal, intake was averaged within each demand-curve price; the resulting per-animal values were then averaged across animals to give group consumption ($\pm$ SEM) at each price. Vehicle sessions ($0 \mu\text{g}/\text{mL}$) were summarized separately and excluded from the fit, as zero-dose consumption is undefined on the logarithmic price axis. Each fentanyl concentration was assigned a standardized unit price C (222.0, 125.0, 70.0, 40.0, and 22.0 $\mu\text{g}/\text{mL}$ mapped to $C = 1, 5, 10, 25, 50$, respectively), following the response-cost schedule and price assignments of Coffey et al. (2023), with the highest dose anchored at $C = 1$ (cost convention of Coffey et al., 2023). Group consumption was fit with the exponentiated demand model (Koffarnus et al., 2015),

$$\log_{10} Q(C) = \log_{10} Q_0 + k \left(e^{-\alpha C} - 1 \right), \quad (3.7)$$

where Q is consumption ($\mu\text{g}/\text{kg}$), Q_0 is demand intensity (consumption as price approaches zero), α is the demand elasticity / essential-value parameter (smaller α = more inelastic demand), and k is the span of consumption in $\log_{10}$ units. Here α and k are the demand-model elasticity and span and are distinct from the pharmacokinetic rate constants α and β above and from the KNN parameter k in the S2C algorithm. Q_0 and α were estimated by nonlinear least squares (`scipy.optimize.curve_fit`, bounded), with k fixed at 2.0 to stabilize the fit given the limited number of price points (Hursh and Roma, 2013); goodness of fit was summarized as R^2 . From the fitted curve we derived the standard behavioral-economics indices: maximal response output ($O_{\max}$) and the price of maximal output ($P_{\max}$) were taken as the maximum and the argmax, respectively, of the expenditure function $Q(C) \cdot C$ evaluated on a dense logarithmically spaced price grid, and essential value was computed as $\text{EV} = 1/(100 \alpha k^{3/2})$.

3.4.16 Cross-session DMS dynamics analysis

Press-aligned dynamics were computed from the event-aligned traces of Fig. 3.5d in a -2 to $+10$ s window around rewarded-press onset. Each neuron’s windowed trace was z -scored to its within-window mean and standard deviation, with the standard deviation floored at 0.5 so near-silent neurons could not dominate the normalization. The Σ Trace is the mean $z \pm$ SEM across all segmented neurons that session; the heatmaps show the 100 neurons with the largest peak $|z|$, ordered by the temporal center of mass of their positive z -activity. Neurons are selected and ordered

independently per session, so rows index population structure rather than tracked-cell identity across concentrations.

For the group-level quantification (Fig. 3.6e), neurons from all animals were pooled per session, the population peak time was taken as the Σ Trace maximum within 0–2 s of the press, and each neuron’s z at that time was summarized as the mean $\pm$ SEM across the pooled population (n neurons); concentrations were compared by Kruskal–Wallis test with trained-vs-each Mann–Whitney U post hoc tests, treating each per-session population as an independent sample.

Single neurons within each concentration were partitioned by the shape of their press-aligned response (Fig. 3.6f, g). Each neuron’s per-neuron z -scored trace (computed as above) was restricted to the 0–10 s post-press window, reduced to its first 10 principal components, and clustered by Ward (variance-minimizing) agglomerative linkage; because the clustering input is per-neuron z , neurons are grouped by response *shape* rather than amplitude, which renders the partition gain-robust and licenses pooling neurons across animals within a concentration. Clusters were ordered by mean z in the 0–2 s press window, so that cluster 1 denotes the most press-excited type. The cluster count was fixed at $k = 2$ on the basis of a per-concentration model-selection sweep over $k = 1$ –7, scored both by Gaussian-mixture Bayesian information criterion (BIC) and by mean silhouette: $k = 1$ was rejected at every concentration ($\Delta\text{BIC} = 511$ –858 for $k=1 \rightarrow 2$), silhouette was maximal at $k = 2$ in every concentration (0.23–0.33) and never exceeded that value for $k \geq 3$, and although BIC continued to decrease weakly for larger k its minimizer was inconsistent across concentrations ($k = 4$ –7) with only small gains ($\Delta\text{BIC} \leq 243$ relative to $k = 2$). Therefore, we implemented the best-separated two-type partition for further analyses. Clustering was performed independently per session on neurons pooled across all six animals and the representative-animal clusters in Fig. 3.6f, g illustrate the typology.

The two-type distinction was operationalized as a single fixed, dose-neutral criterion rather than the per-session Ward labels. A neuron was classed press-responsive if the peak of its per-neuron z -scored trace within 0–3 s of the rewarded press exceeded 2, and non-responsive otherwise. We took the fraction of segmented neurons in each class, summarized these as the across-animal mean $\pm$ SEM, and compared classes across concentration with a Friedman test (Kendall’s W effect size), treating animals as repeated-measures blocks; this same fixed criterion defines the response classes carried into the cross-session tracking analyses (Fig. 3.6h).

For the session-to-session flow analysis (Fig. 3.6i), we assigned the neurons that S2C registered across both sessions a given pair (222 $\rightarrow$ 70, 70 $\rightarrow$ 22, 22 $\rightarrow$ 0), and assigned each to press-responsive or non-responsive at each concentration by the same fixed criterion used for panel h. Each concentration column is normalized to unit height so that band heights are class fractions and each ribbon is drawn with width proportional to the number of co-tracked neurons making a given class-to-class transition. Transitions were summarized as the fraction of pair-tracked neurons that changed class between adjacent concentrations.

For the cross-session responsive-class analyses (Fig. 3.6j), each neuron S2C-registered across a session pair was labeled press-responsive or non-responsive at each session by the fixed $z > 2$ criterion

above. Agreement between the two labels was quantified by the adjusted Rand index (ARI) over the cells co-tracked in that pair, with 95% confidence intervals from 2,000 cell-level bootstrap resamples. Turnover was defined as the fraction of co-tracked cells changing class, and retention the fraction of session- s responsive cells still responsive at session- s' , compared against the marginal responsive rate (chance). To separate biological reorganization from registration error we computed a registration ceiling expected if responsive identity were perfectly stable and the only disagreements arose from S2C mismatches. This simulation was performed by assigning each cell its reference-session label and then each session’s label was independently flipped with probability $1 - F1 \approx 0.015$. The panel *j* raster visualizes this contrast, with the “Theoretical Stability” raster generated by the same procedure and the “Observed” raster showing the measured labels, both restricted to cells tracked across all four shown sessions and sorted by first-session status. As a stimulus-held-fixed control, the ARI and turnover analyses were repeated on the constant-concentration ($70 \mu\text{g/mL}$) chain of consecutive sessions. To confirm the lability reflected the press response rather than the binary threshold, we recomputed cross-session agreement on a noise-robust graded-response measure via the baseline-subtracted mean of the press-window trace. This computation carried no cross-session structure specific to the press window, confirming the turnover is a property of the population and not an artifact of the binary criterion.

3.4.17 Statistical analysis

All analyses in Python 3.10 (SciPy, statsmodels). For the S2C-vs-ROI-matching comparison, the unit of paired analysis was a single F1 score per (animal, seed) within each (condition, neuron-count) cell, computed by averaging F1 across the four consecutive session transitions of that animal-seed’s 5-session chain. Each such F1 was paired across tools (S2C vs. CellReg) on identical underlying centroid data and ground-truth identity. The full paired-run count therefore arrives at 32 productive (condition $\times$ neuron-count) cells $\times$ 8 animals $\times$ 5 seeds = 1,280 paired comparisons in principle per footprint scenario. A paired comparison was excluded only when CellReg returned no defined F1 for any of the four consecutive transitions (Search-radius choice); with no per-pair F1 to pair against the corresponding S2C value, those runs cannot enter the paired statistic without imputation, which we declined to perform. This yields $n_{\text{grand pooled}} = 1,262$ in the worst-case scenario (18 dropped) and $n_{\text{grand pooled}} = 1,262$ in the best-case scenario (18 dropped). We note that this 1,262 is the unit of *statistical comparison*, not the count of S2C invocations: each chain involved 5 separate S2C runs (one per session) and produced 4 session-pair F1 values, so the underlying number of individual S2C runs is $5\times$ larger ($\sim 6,310$) and the underlying number of session-pair F1 measurements is $4\times$ larger ($\sim 5,048$). Per-condition tests used the paired Wilcoxon signed-rank statistic with Holm-Bonferroni correction across the 8 conditions within each neuron-count tier; effect sizes were reported as paired Cohen’s d_z ; 95% confidence intervals on the F1 difference were obtained via bias-corrected bootstrap (10,000 resamples). Pooled tests across conditions and neuron counts used the same paired-test framework on the full pooled set. For the temporal-correlation control, mean accuracy per duration was tested against chance ($1/N = 0.10\%$) by paired t -test across the 10 session pairs. $\alpha = 0.05$

throughout. Data are presented as mean $\pm$ SEM unless otherwise noted.

Data availability

The processed data supporting the findings of this study are deposited on Zenodo and released under CC-BY-4.0 (Peden-Asarch et al., 2026b). This deposit is comprised of the MPS-extracted spatial footprints + traces, behavioral event tables, and S2C track assignments. The benchmark is reproducible from the released generator (Code availability). Raw single-photon miniscope videos (~1.4 TB) are available from the corresponding author on reasonable request; they are not deposited due to their size and are not required to interpret, verify, or extend the reported results.

Code availability

Stars2Cells is distributed as a GUI-driven standalone application for macOS (a signed and notarized DMG containing Stars2Cells.app) and Windows (a ZIP archive containing a standalone Stars2Cells.exe); both builds bundle a self-contained Python runtime, so users do not need to install Python or manage virtual environments. Source code is available at https://github.com/ariasarch/Stars2Cells_GUI, and the exact release used to produce the results in this manuscript is archived on Zenodo (Peden-Asarch et al., 2026a). All other custom code used to generate the results reported here is collected in a single companion repository (https://github.com/ariasarch/S2C_COMPANION_REPO). All code is released under the MIT license.

Acknowledgments

This work was supported by the National Institute on Drug Abuse (R01DA052618 and R00DA052571), a VA Puget Sound Health Care System Research & Development Pilot award, National Institute on Drug Abuse training grant T32DA007278.

Author contributions

A.P.-A. Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing – original draft; Writing – review & editing; Visualization. **L.H.** Data curation (DMS dataset processing through MPS). **J.B.** Investigation (movement/kinematic analysis). **E.A.** Software (Astropy porting); Validation (astrometric testing). **J.Q.** Investigation (DMS behavioral-economics fentanyl self-administration experiments, including animal training and histological verification). **K.R.C.** Conceptualization; Resources; Writing – review & editing. **J.F.N.** Conceptualization; Supervision; Funding acquisition; Resources; Project administration; Writing – review & editing.

Competing interests

The authors declare no competing interests.

Appendix B

Supplementary material for Chapter 3

Supplementary Note 1: Ownership Guard Uniqueness

Proposition B.1 (Unique Emission). *Let $Q = \{P_a, P_b, P_c, P_d\}$ be a quad with six pairwise distances. Under the ownership guard, Q is emitted exactly once across the entire diagonal set: from the single diagonal equal to its longest pairwise edge.*

Proof. Define the ownership diagonal of Q as the pair (i^*, j^*) achieving the maximum pairwise distance:

$$(i^*, j^*) = \arg \max_{(i,j) \in \binom{Q}{2}} |P_i - P_j| \quad (\text{B.1})$$

When processing diagonal (d_1, d_2) , the ownership guard emits Q only if:

$$\max_{(i,j) \in \binom{Q}{2}} |P_i - P_j| \leq |P_{d_2} - P_{d_1}| \cdot (1 + \epsilon), \quad \epsilon = 10^{-6} \quad (\text{B.2})$$

This condition is satisfied if and only if (d_1, d_2) is the longest edge (within floating-point tolerance). For generic point configurations (no two pairwise distances exactly equal), exactly one of the six edges is longest, so Q is emitted exactly once. Exact ties occur with probability zero for real-valued centroid coordinates. $\square$

Complexity reduction. Without the ownership guard, each quad could be emitted from up to 6 diagonals (one per pairwise edge that appears in the sampled diagonal set), requiring an $O(Q \log Q)$ global deduplication pass via `np.unique`. With the guard, the output is duplicate-free by construction and the deduplication pass is eliminated entirely.

Supplementary Note 2: $\sqrt{N}$ Threshold Scaling

The optimal descriptor similarity threshold τ must balance two forces: too low and same-cell matches are rejected (low recall); too high and false matches dominate (low precision). The per-animal

calibration in Step 1.5 finds the optimum τ at each N by sweeping thresholds on sampled quad subsets, and the empirical scaling across the four neuron-count tiers tested is well-fit by $\tau = C\sqrt{N}$. Below we derive this scaling from the mean-KNN-distance behavior of the local-diagonal class, and show why the alternative scaling implied by 4-D descriptor density is non-binding in the operating scenario.

Setup. S2C builds quads from two diagonal classes: (i) KNN-local diagonals connecting each neuron to its k nearest neighbors, and (ii) random long-range diagonals spanning the FOV. The two classes have very different diagonal-length scaling with N and, consequently, very different descriptor-blur behavior. The threshold is set by whichever class is worse-blurred, which turns out to be the local class.

Step 1: local diagonal length scales as $N^{-1/2}$. For N points distributed in a fixed FOV of area A , the mean distance to the k -th nearest neighbor scales as

$$L_q^{\text{local}} \sim \sqrt{kA/N} \propto N^{-1/2} \quad (\text{fixed } k, A). \quad (\text{B.3})$$

This holds for both Poisson and Poisson-disk-sampled point patterns; only the prefactor differs (Poisson-disk is more concentrated around its mean than Poisson, but both inherit the $N^{-1/2}$ exponent from constant-density scaling). The intuition is direct: in a denser field, your nearest neighbor is simply closer.

Step 2: same-cell descriptor blur scales as $\sqrt{N}$. Per-neuron centroid jitter of standard deviation σ_j propagates to descriptor space as $\sigma_d \approx \sqrt{2} \sigma_j / L_q$ (Methods). Substituting Step 1:

$$\sigma_d^{\text{local}} \approx \frac{\sqrt{2} \sigma_j}{\sqrt{kA/N}} = \sigma_j \sqrt{\frac{2N}{kA}} \propto \sqrt{N}. \quad (\text{B.4})$$

The intuition: at high density the same absolute jitter occupies a larger fraction of the (now shorter) diagonal length, and that ratio is exactly what propagates into descriptor space.

Step 3: the threshold tracks same-cell blur. A same-cell quad pair has descriptor distance $\sim \sigma_d$ in expectation. For same-cell matches to survive descriptor filtering, τ must exceed σ_d by a constant capture margin $\beta > 1$:

$$\tau \gtrsim \beta \sigma_d^{\text{local}} = \underbrace{\beta \sigma_j \sqrt{\frac{2}{kA}}}_C \sqrt{N}. \quad (\text{B.5})$$

The constant C absorbs the capture margin β (set by the desired same-cell capture fraction; e.g., $\beta \approx 2$ captures $\sim 95\%$ of a Gaussian), the per-animal jitter level σ_j , the usable FOV area A , the KNN parameter k , and an $O(1)$ geometric prefactor relating mean KNN distance to typical local-quad diagonal length. None of these depend on N , so the leading-order behavior is exactly $\tau \propto \sqrt{N}$. The empirical per-animal calibration of C recovers exactly this group of experimental constants.

The 4-D descriptor-density floor is non-binding An alternative starting point is descriptor-space density: $Q \propto N$ quads occupy a compact 4-D region, so the expected nearest-neighbor distance

among descriptors scales as $NN_d \propto N^{-1/4}$ and the false-match floor in cosine distance shrinks accordingly. This argument predicts τ should also *shrink* with N – the opposite direction from both the empirical fit and Step 3 above. The reconciliation is that same-cell blur grows with N much faster than the false-match floor shrinks; their ratio is

$$\frac{\sigma_d^{\text{local}}}{NN_d} \propto \frac{N^{1/2}}{N^{-1/4}} = N^{3/4}. \quad (\text{B.6})$$

Throughout the operating range, same-cell blur is the binding constraint by a margin that widens with N ; the false-match floor is non-binding. False matches that survive the descriptor stage are filtered downstream by the consistency check and RANSAC, which are explicitly designed to absorb high false-match fractions.

Why long-range diagonals are insensitive to N . Random long-range diagonals span the FOV regardless of density, $L_q^{\text{long}} \sim \sqrt{A}$, so their descriptor blur $\sigma_d^{\text{long}} \approx \sqrt{2} \sigma_j / \sqrt{A}$ is independent of N . This is the design reason both classes exist. Local diagonals stabilize matching under small inter-session shifts but set the threshold; long-range diagonals carry the matching signal at high density and operate well below τ . The two-class construction is what makes per-animal calibration tractable: C only needs to track local-diagonal blur, while the bulk of useful matches come from the long-range class.

Supplementary Note 3: Coverage Remediation

The ownership guard (Supplementary Note 1) eliminates duplicate quads but can cause neurons in dense local regions to have zero or very few quads. This occurs when a neuron only participates in short diagonals, all of which have their quads claimed by longer neighboring diagonals.

Detection. For each neuron i , define coverage κ_i as the number of quads containing i . The field median coverage is $\tilde{\kappa} = \text{median}(\{\kappa_i\}_{i=1}^N)$. Neuron i is *undercovered* if:

$$\kappa_i < f_{\min} \cdot \tilde{\kappa} \quad (\text{B.7})$$

where $f_{\min} = 0.4$ by default.

Remediation. For each undercovered neuron i , all diagonals (i, j) and (j, i) in the diagonal bucket dictionary are re-processed by `__process_diagonal` with `enforce_ownership=False`. The resulting quads are subject to all other quality filters (area ≥ 1.0 , min pairwise distance, non-degeneracy) and deduplicated against the existing quad set via a Python set of sorted index tuples. Quality pruning (quad_keep_fraction) is applied to the remediation batch independently.

Guarantees. Remediation quads may duplicate quads that would have been emitted from a different diagonal under normal ownership. However, because they are deduplicated against the existing set, no quad appears twice in the final output. The trade-off is that these remediation quads lose the automatic "emitted exactly once" guarantee the ownership guard gives the main pass: because they are produced by diagonals that are not their rightful owner, their uniqueness is

enforced by the explicit duplicate check rather than guaranteed by construction.

Supplementary Note 4: Computational Complexity

Step	Operation	Complexity
1 (Diag. buckets)	KD-tree + edge list	$O(N \log N + N \cdot k)$
1 (Heights)	Per-diagonal height computation	$O(\mathcal{E} \cdot N)$
1 (Quads)	Pair third points + ownership	$O(\mathcal{E} \cdot K^2)$
1.5 (Calibration)	Threshold sweep on subsets	$O(n_{\text{pairs}} \cdot n_{\tau} \cdot Q_{\text{sample}})$
2 (Matching)	Cosine NN search	$O(Q_r \cdot Q_t)$ or $O(Q_r \log Q_t)$ with FAISS
2.5 (RANSAC)	Batched SVD + chunked residuals	$O(K_{\text{iter}} \cdot m^2 + K_{\text{iter}} \cdot M/C)$
3 (Hungarian)	Vote matrix + assignment	$O(M \cdot 16 + n^3)$
3 (Consolidation)	Transitive chaining	$O(S \cdot n)$

where N = neurons, $k = k_{\text{local}} + k_{\text{random}}$, $|\mathcal{E}|$ = number of diagonals $\approx Nk/2$, K = top- K cap per diagonal, Q = total quads, M = matched quads, $n = \min(N_r, N_t)$, S = number of sessions.

Supplementary Note 5: Synthetic Degradation Model

Base field generation. For each synthetic animal, a base centroid field $\{(x_i^{(0)}, y_i^{(0)})\}_{i=1}^N$ is placed in a 600×600 -pixel FOV by Bridson Poisson-disk sampling, producing a blue-noise (well-separated, non-clustered) layout. The minimum spacing r_{min} defaults to 55% of the theoretical hex-packing maximum for the requested N within the usable FOV (margin = 10 px), so density automatically scales with neuron count. A fallback rejection-sampling step with relaxed r_{min} (factor 0.8) handles edge cases where Bridson cannot place all requested points in the first pass.

Per-neuron base jitter. Independent of session-to-session perturbations, each neuron carries a stable per-neuron centroid offset $(\delta x_i, \delta y_i) \sim \mathcal{N}(0, \sigma_b^2)$ with $\sigma_b = 0.5$ px, modeling systematic CNMF centroid bias that persists across sessions for that neuron. This is added in addition to a per-session perturbation $\sigma_p = 0.2$ px drawn fresh each session.

Session generation. For session $s \in \{1, \dots, S\}$ (where $S = 5$), the base field is transformed in the following fixed order:

1. **Rotation:** $\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = R(\theta_s) \begin{pmatrix} x_i^{(0)} - \bar{x} \\ y_i^{(0)} - \bar{y} \end{pmatrix} + \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$, where $R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$, with θ_s accumulating across sessions in Tiers A and B and following the random-walk schedule in Tier C.
2. **Translation:** $(x'_i, y'_i) \leftarrow (x'_i + t_x^{(s)}, y'_i + t_y^{(s)})$, similarly accumulating per tier.

3. **Clip-to-FOV:** any neuron whose transformed position falls outside $[0, 600]^2$ is removed, modeling loss off the sensor edge after large transforms (this is what makes large rotation/translation conditions genuinely hard rather than merely shifted).
4. **Dropout (sessions ≥ 2 only):** a fraction δ_s of remaining neurons is randomly removed, modeling CNMF detection failure. *No new neurons are spawned* – this matches the biological fact that ROIs can fail to be detected in a given session, but new cells do not spontaneously appear.
5. **Jitter:** $(x'_i, y'_i) \leftarrow (x'_i + \delta x_i + \epsilon_x^{(s)}, y'_i + \delta y_i + \epsilon_y^{(s)})$, where $(\delta x_i, \delta y_i)$ is the stable per-neuron base offset and $\epsilon_x^{(s)}, \epsilon_y^{(s)} \sim \mathcal{N}(0, \sigma_p^2)$ is the per-session perturbation, applied only to neurons that survived clipping and dropout.
6. **Ground-truth capture:** the surviving base indices (before permutation) are stored as ground `_truth_base_ids` in the per-file ground-truth dictionary.
7. **ID permutation:** ROI identifiers are randomly shuffled to prevent matching by index.

Supplementary Note 6: CellReg Reimplementation and Validation

ROI-based matching via a probabilistic model. CellReg (Sheintuch et al., 2017b) models the probability that neurons i (session 1) and j (session 2) are the same cell as:

$$P(\text{same} \mid \rho_{ij}, d_{ij}) = \frac{P(\rho_{ij}, d_{ij} \mid \text{same}) \cdot P(\text{same})}{P(\rho_{ij}, d_{ij})} \quad (\text{B.8})$$

where ρ_{ij} is the spatial footprint correlation and d_{ij} is the centroid distance. The likelihoods are modeled as:

$$P(\rho, d \mid \text{same}) = f_{\text{same}}(\rho) \cdot g_{\text{same}}(d) \quad (\text{B.9})$$

$$P(\rho, d \mid \text{diff}) = f_{\text{diff}}(\rho) \cdot g_{\text{diff}}(d) \quad (\text{B.10})$$

where $g_{\text{same}}(d)$ is modeled as a Rayleigh distribution (centroid distances for matched neurons) and $g_{\text{diff}}(d)$ as approximately uniform over the FOV. The correlation distributions $f_{\text{same}}(\rho)$ and $f_{\text{diff}}(\rho)$ are estimated non-parametrically from the data. Parameters are fit via expectation-maximization over the observed pairs.

Python reimplementation. The step-by-step port of CellReg’s spatial-correlation pathway, its `cluster_cells` final-assignment step (which differs from a one-shot Hungarian solve), and its pair-for-pair validation against native MATLAB CellReg are given in Methods (“ROI-based matching reimplementation”). The probabilistic model above is what those steps implement.

Parallelization. The validated reimplementation was wrapped in an embarrassingly-parallel batch driver (Python’s `concurrent.futures.ProcessPoolExecutor`, up to 50 workers, each pinned to a

single BLAS thread) so the full benchmark runs on one workstation; the algorithm is unchanged and only its execution is parallelized.

Supplementary Note 7: Related tools and why they are not head-to-head comparators

SCOUT builds on CellReg’s spatial-correlation backbone, so CellReg upper-bounds the part of SCOUT that S2C actually competes with; SCOUT’s added temporal-correlation cue is the chance-level signal we characterize directly (Section 3.2.2, Fig. 3.3), and its remaining additions (Jensen–Shannon divergence on elliptic-Gaussian footprint fits, metric-weight perturbation, and consensus clustering) are footprint-level structural cues that require pixel footprints our centroid-only benchmark does not provide. Those are a separate, footprint-level question, not a different answer to ours. Centroid-only predecessors also exist like STAT (Gu et al., 2023) and Coherent Point Drift (Myronenko and Song, 2010), but these recover identity by fitting a single spatially-smooth deformation between sessions, an assumption that degrades precisely under the miniscope recording conditions that S2C’s transform-free invariant descriptors absorb by construction. Benchmarking any of these orthogonal methods against S2C would score them on a task they were never built for.

Supplementary Note 8: Full Per-Condition Benchmark Statistics

Table B.1 expands the pooled and per-neuron-count summaries in the main text (Fig. 3.4g) to all 32 (condition $\times$ neuron-count) benchmark cells. S2C’s F1 advantage over ROI-based matching is positive and significant in every cell, with the smallest gaps on the mildest perturbation (Tier-A translation) and the largest on the high-displacement Tier-B/C conditions.

Table B.1: Per-condition S2C vs. ROI-based matching (worst-case footprint scenario). Mean F1 per tool, the paired S2C–ROI F1 difference with 95% bootstrap CI, paired Cohen’s d_z , and Wilcoxon p (Holm-corrected across conditions within each neuron-count tier). $n = 40$ paired runs per cell except where CellReg returned no defined F1 (excluded per Methods); S2C precision was $\geq 97.7\%$ throughout; best-case footprint values are near-identical.

Condition	n	S2C F1 (%)	ROI F1 (%)	Δ F1 (pp) [95% CI]	d_z	p_{Holm}
<i>N = 100 neurons</i>						
A: rotation	40	99.6	76.0	+23.5 [+15.2, +32.4]	+0.82	2.3×10^{-9}
A: translation	40	99.4	97.8	+1.6 [+0.8, +2.6]	+0.57	9.4×10^{-4}
A: combined	40	99.4	68.8	+30.6 [+22.0, +39.3]	+1.10	1.5×10^{-11}
B: dropout	40	99.9	47.9	+52.0 [+46.0, +57.6]	+2.68	1.5×10^{-11}
B: drift	35	99.8	19.3	+80.6 [+74.9, +85.9]	+4.84	1.7×10^{-10}
B: rotation	39	99.9	14.4	+85.5 [+79.9, +90.1]	+5.15	1.8×10^{-11}
B: combined	39	99.8	12.3	+87.5 [+83.2, +91.2]	+6.75	1.8×10^{-11}
C: random walk	40	99.7	23.1	+76.6 [+70.9, +82.0]	+4.17	1.5×10^{-11}
<i>N = 250 neurons</i>						
A: rotation	40	99.9	69.0	+30.9 [+21.5, +41.0]	+0.98	1.5×10^{-11}
A: translation	40	99.9	97.4	+2.5 [+1.2, +4.0]	+0.54	6.3×10^{-7}
A: combined	40	99.9	62.0	+37.9 [+28.7, +47.0]	+1.24	1.5×10^{-11}
B: dropout	40	99.9	36.8	+63.1 [+59.1, +67.1]	+4.81	1.5×10^{-11}
B: drift	38	99.9	15.4	+84.5 [+78.7, +89.9]	+4.86	1.6×10^{-7}
B: rotation	40	99.9	7.6	+92.4 [+89.9, +94.5]	+12.27	1.5×10^{-11}
B: combined	39	99.9	6.8	+93.1 [+91.2, +94.8]	+15.58	1.6×10^{-7}
C: random walk	40	99.9	20.1	+79.8 [+76.5, +82.8]	+7.63	1.5×10^{-11}
<i>N = 500 neurons</i>						
A: rotation	40	97.2	54.7	+42.4 [+33.4, +51.3]	+1.45	1.6×10^{-11}
A: translation	40	97.4	92.9	+4.5 [+2.3, +6.9]	+0.61	4.6×10^{-3}
A: combined	40	97.2	45.0	+52.2 [+44.2, +59.7]	+2.07	1.5×10^{-11}
B: dropout	40	97.0	30.5	+66.5 [+63.9, +69.0]	+7.86	1.5×10^{-11}
B: drift	35	97.0	11.3	+85.6 [+81.5, +89.5]	+6.97	1.2×10^{-10}
B: rotation	40	96.7	6.3	+90.4 [+88.1, +92.4]	+12.78	1.5×10^{-11}
B: combined	39	96.8	5.2	+91.6 [+90.2, +92.9]	+21.03	1.5×10^{-11}
C: random walk	40	96.7	15.4	+81.3 [+78.5, +83.9]	+9.19	1.5×10^{-11}
<i>N = 1,000 neurons</i>						
A: rotation	40	94.3	39.8	+54.4 [+45.9, +62.7]	+1.97	1.5×10^{-11}
A: translation	40	94.9	84.3	+10.6 [+6.8, +14.6]	+0.82	2.6×10^{-4}
A: combined	40	94.2	32.7	+61.5 [+55.0, +67.4]	+3.07	1.5×10^{-11}
B: dropout	40	99.3	23.2	+76.1 [+75.0, +77.0]	+22.98	1.5×10^{-11}
B: drift	38	100.0	7.9	+92.1 [+89.2, +94.8]	+10.12	2.2×10^{-11}
B: rotation	40	99.8	4.4	+95.3 [+94.0, +96.6]	+22.20	1.5×10^{-11}
B: combined	40	97.4	3.4	+93.9 [+92.0, +95.6]	+15.70	7.1×10^{-8}
C: random walk	40	98.4	10.5	+87.9 [+85.9, +89.8]	+13.90	1.5×10^{-11}

Chapter 4

The Lateral Habenula Retrodictively, Not Predictively, Reports the Outcome of Adverse Consequences Under Fentanyl Self-Administration

In preparation as: Peden-Asarch AM, Sievertsen SA, Sanders J, Lin C-Y, Belur N, Nickelson WB, Neumaier JF, Coffey KR. JFN and KRC are co-senior authors.

4.1 Introduction

The opioid crisis continues to devastate communities worldwide, with over 105,000 overdose deaths in the United States in 2023 (Garnett and Miniño, 2024). Synthetic opioids, primarily illegally manufactured fentanyl, account for the large majority of opioid-related deaths, marking the “third wave” of the epidemic (Volkow and Blanco, 2021). Understanding the neural mechanisms governing the decisions during active drug onboarding remains a critical challenge for developing effective interventions.

Traditional models of addiction and relapse emphasize decisions made during abstinence, where the incubation-of-craving model proposes that negative drug-associated outcomes lose influence over time and craving re-emerges to drive relapse (Pickens et al., 2011; Venniro et al., 2016; Chow et al., 2025; Shalev et al., 2002). Complementing this body of work, another common feature of human substance use disorder is that continued use while the drug is already onboard (McNally et al., 2023). After consuming an initial dose and experiencing its pharmacological effects, individuals encounter repeated points in which they must decide whether to continue use despite awareness of some potential negative consequences. Decision-making under acute drug exposure places distinct demands on neural systems, as valuation, affect, and action selection occur in the presence of

ongoing pharmacological modulation. Thus, models that test relapse during abstinence do not capture a defining feature of substance use disorder, in which self-administration continues beyond what the individual intended, and this may reflect an alteration in valuation while the drug is onboard. While several preclinical paradigms have examined shocked self-administration and voluntary abstinence (Fredriksson et al., 2021b; Venniro et al., 2020; Marchant et al., 2013; Pelloux et al., 2007; Pascoli et al., 2023), comparatively less is known about how neural circuits operate during these within-session decisions made under sustained drug exposure.

The lateral habenula (LHb) plays a key role in integrating information about negative events. For example, LHb neurons fire when an outcome is worse than expected and go quiet when an outcome is better than expected (Matsumoto and Hikosaka, 2007, 2009). That signal is relayed through the rostromedial tegmental nucleus onto midbrain dopamine neurons, which the LHb inhibits, and onto serotonergic raphe nuclei (Jhou et al., 2009; Hong et al., 2011; Proulx et al., 2014). The LHb is therefore positioned to contribute to behavioral flexibility when negative consequences mount.

One way to assess LHb contribution to decision making is to vary the reinforcement value, predictability, or to introduce a cost to self-administration. These questions have been taken up by two largely separate literatures, one asking what the LHb does during self administration (the drug axis) and one asking what the LHb does during stress and fear (the shock axis). On the drug axis, LHb activity has been shown to track the value of the drug itself, rising to drug-predictive cues and dipping during consumption (Coffey et al., 2023). On the shock axis, LHb neurons respond to externally imposed aversive stimuli including shock, quinine, and air puff, with most excited and a minority inhibited (Matsumoto and Hikosaka, 2009; Wang et al., 2017; Congiu et al., 2019; Shabel et al., 2019), LHb inactivation leaves the suppression of shocked instrumental responding intact and shifts only response latency (Jean-Richard-dit Bressel and McNally, 2014), and habenular circuit engagement covaries with whether an animal keeps responding once its self-administration is shocked (Duan et al., 2022). Arguably, neither arrangement reproduces the state that defines opioid use disorder, in which the choice to continue is made while the drug is pharmacologically active and while the potential consequence of continuing is known.

The intensity of the shock is itself a design parameter, and the two literatures settled on different values for it. Shocked self-administration settled on mild intensities near 0.2 to 0.5 mA (Engeln and Ahmed, 2024), a convention inherited from the operant tradition (Azrin and Holz, 1966) and carried into addiction research by the seminal shocked cocaine studies (Deroche-Gamonet et al., 2004; Pelloux et al., 2007; Jonkman et al., 2012). The reasoning being that a mild shock leaves a persistent subgroup that still responds despite the shock (Marchant et al., 2018), whereas raising the intensity to 0.8 mA or above compresses the individual differences being measured leaving no group responding (Engeln and Ahmed, 2024). On the other hand, stress and fear work settled on 0.8 mA instead, a value that traces to the rodent learned helplessness protocol (Vollmayr and Henn, 2001) and entered the LHb field through helplessness models of depression (Li et al., 2011; Yang et al., 2018). Acute recordings of LHb shock encoding have used 0.3 mA (Lecca et al., 2017). Thus, no study has recorded the same LHb neurons across a range of intensities, and whether the behavioral

break between mild and suppressive shock corresponds to a break in LHb encoding is untested.

Fentanyl self-administration is not uniform within a session, since pharmacokinetic modeling of operant response patterns indicates that an animal responds rapidly to reach a preferred brain concentration of drug (loading) and then responds more slowly to maintain that level (maintenance) (Coffey et al., 2023). This loading and maintenance pattern is itself a strategy of self administration, and introducing an adverse consequence during it requires the animal to change that strategy while the drug is onboard. Read as a session response rate, that change is averaged away along with the trial on which it occurred. Resolving when an animal changes strategy therefore requires a within-session changepoint rather than a session total, and it also supplies the trial at which to ask what the LHb was doing. Moreover, since ultrasonic vocalizations index affective state independently of responding (Brudzynski, 2013; Coffey et al., 2019), the affective consequence of stopping can be measured rather than inferred from stopping itself. Given that the LHb signals worse-than-expected outcomes, an adverse consequence delivered while the drug is onboard is one such outcome, and whether its activity runs ahead of the strategy change or behind it is what separates computing the change from reporting it.

If different LHb neurons encode both drug value and shock, then the averaged signal across many neurons (such as with fiber photometry) may cancel each other out. Recovering both signals in the same population therefore requires placing the shock inside the window of active drug consumption, which in turn requires a delay between the press and the shock. Moreover, most *in vivo* work in this structure has used fiber photometry collapsing every neuron into one trace, which resolves a population only to the extent that its members respond alike (Legaria et al., 2022). Therefore, a heterogeneous population within a brain area like the midbrain and It is reasonable to induce that the LHb contains this kind of heterogeneity as evidenced by: 1) LHb glutamatergic afferents arrive from segregated sources including medial prefrontal cortex, ventral pallidum, lateral hypothalamus, lateral preoptic area, and the entopeduncular nucleus, each carrying different information about cognitive control, incentive value, homeostatic state, and nociception (Hu et al., 2020; Hikosaka, 2010; Barker et al., 2017), 2) LHb efferents are split between dopaminergic and serotonergic targets (Hikosaka et al., 2008), and 3) Multiple molecular subtypes coexist within the nucleus (Hashikawa et al., 2020; Wallace et al., 2020). Thus, since an average cannot test that heterogeneity given that it is unable to separate multiple or opposing signals, then it cannot answer whether the neuron carrying a signal this week carried it last week.

One solution to the dilemma of bulk activity from divergent signals canceling out is to preserve the signals of individual neurons. Miniaturized microscopes (miniscopes) allow chronic calcium imaging of hundreds of neurons in freely behaving animals across multiple sessions (Ghosh et al., 2011a; Cai et al., 2016b), providing cellular-resolution access to population dynamics. However, existing processing pipelines struggle with the computational demands of multi-hour recordings typical of addiction paradigms. Here we utilized our recently released software - Miniscope Processing Suite (MPS; (Peden-Asarch et al., 2025)) to solve this limitation by providing a scalable pipeline optimized for long-duration sessions. Holding neuronal identity across sessions separated by days

poses a second problem since the field of view shifts between mountings and the scope itself may be remounted in a different orientation. We used Stars2Cells, which adapts astrometric plate solving to this registration problem, to track the same neurons across the escalation ladder and Relapse. Together these tools give cellular-resolution access to the same population across weeks, which is what separates a signal that is absent from one that is averaged away.

Here we developed a delayed shock paradigm for oral fentanyl self-administration in rats. After a baseline of drug onboarding was established over weeks, a shock was delivered 4 s after each rewarded press, following an initial loading period that leaves fentanyl pharmacologically active across the decision window. Shock intensity escalated across days from 0.1 to 0.8 mA, spanning both conventions, followed by a Relapse test one week later. We combined chronic miniscope imaging of LHb neurons processed with the Miniscope Processing Suite (MPS) (Peden-Asarch et al., 2025), cross-session registration of the same neurons with Stars2Cells (S2C) (Peden-Asarch et al., 2026d), a within-session changepoint model to localize each animal’s shift in strategy, ultrasonic vocalization recording to index affective state via DeepSqueak (Coffey et al., 2019), and automated pose tracking to rule out gross motor impairment. This design asks whether the LHb signals an adverse consequence during a shift in strategy of fentanyl self-administration while the drug is still on board, whether that signal is predictive or retrodictive of the shift, whether it survives over weeks, and whether the same neurons carry it.

4.2 Methods

4.2.1 Animals and housing

For all experiments, 8-week-old male and female Long-Evans rats were bred in-house. Rats were dual-housed in a temperature and humidity-controlled room on a 10:14 dark-light cycle and acclimated to the facility for ~ 1 week before beginning experimental procedures at 9-10 weeks old. Food and water were available ad-lib throughout the experiment (weight range: 250-450g). Experimental procedures were approved by the University of Washington Institutional Animal Care and Use Committee.

4.2.2 Oral Fentanyl Self-Administration

We modified our previously published procedure (Coffey et al., 2023) as follows. Male ($n=19$) and female ($n=14$) rats were first trained to self-administer 0.1% saccharin solution alone in a 1-h session under the same lever-press contingency that would later deliver oral fentanyl. Of these, 6 were run in an earlier pilot cohort that did not include the delayed shock paradigm and contributed to acquisition and baseline measures only, so the escalation, Danger-phase, and Relapse analyses used the remaining 27 animals. After this initial training day, rats responded for liquid fentanyl (70 $\mu\text{g/mL}$ in H_2O , 0.05 mL/delivery) five days per week in 3-hour sessions in standard operant chambers (Med Associates) equipped with two levers, reward delivery port, white stimulus light, house light, and tone generator. Fentanyl was paired with a descending saccharin gradient (0.1%, 0.08%, 0.06%,

0.04%, 0.02%) with a 70 $\mu\text{g}/\text{mL}$ fentanyl dose present from Day 1 to Day 5. Including the saccharin fade, from Days 1-10 the animals underwent a standard fixed-ratio (FR-1) self-administration where each active lever press resulted in: (1) 0.05 mL fentanyl delivery, (2) white stimulus light illumination over the lever light (10 s), (3) 4-kHz tone presentation (10 s), and (4) house light extinction (10 s). A 10 s timeout was imposed during the CS presentation. Inactive lever presses were recorded. Sessions terminated after 3 hours. For Days 11-19, sessions transitioned to a trial-based structure with 180 total trials per session. For days 11-15, the trial window was increased from 10 s to 15 s, and then to 20 s thereafter to accommodate miniscope operation (5 s pre-trial and 5 s post-trial), allowing the LED to be switched off during the timeout while still capturing activity immediately before and after each trial.

Beginning on Day 20, sessions included two distinct phases. Safe loading (trials 1-15), identical to the training phase, allowing drug loading to achieve preferred brain concentrations (Fig. 4.1b). Danger phase (trials 16-180) where following each drug delivery, a 4 s delay period was imposed, then a 1 second shock was delivered through the grid floor (Fig. 4.1c). During the Danger phase, the houselight was switched from white to green and repositioned to signal that lever pressing would result in shock delivery, and the lever-associated cue light no longer illuminated upon an active press and a higher tone (+1 kHz) was emitted, ensuring animals could discriminate Danger trials from Safe trials via altered cue structure. Animals could complete all total trials per session. Shock escalation schedule: Day 20 (0.1 mA), Day 21 (0.2 mA), Day 22 (0.4 mA), Day 23 (0.8 mA), Day 30 (0.8 mA, Relapse). After completing the Danger sessions, animals underwent one week of forced abstinence and were then returned to the operant chamber on Day 30, where the full trial structure was reinstated exactly as before, with drug still available to assess Relapse-like responding. Each trial spanned from the active lever press to the end of the inter-trial interval (20 s post-shock or post-timeout). Miniscope recordings occurred on Days 19, 21, 23, and 30 to capture baseline, early negative consequence, strong suppression, and persistent abstinence states while preserving equipment integrity. Behavioral metrics: Total rewarded lever presses, head entries, latency (lever press to head entry), active/inactive lever presses, estimated fentanyl intake ($\mu\text{g}/\text{kg}$ based on body weight and rewarded lever presses count) were recorded.

For each rewarded lever press, the dose in micrograms per kilogram was calculated as fentanyl concentration (70 $\mu\text{g}/\text{mL}$) multiplied by infusion volume (0.05 mL) and divided by the animal's body weight in kilograms. Whole-brain fentanyl concentration over time was then estimated using the same two-compartment model applied to this oral fentanyl paradigm (Coffey et al., 2023) based on (Pan et al., 1991). Briefly, each infusion contributed a bi-exponential term with a fast distribution component and a slower elimination component, parameterized by a transfer constant k , an apparent brain volume v , and rate constants α and β . In practice, this was implemented as a convolution of the infusion times with a kernel proportional to $\exp(-\alpha \cdot \Delta t) - \exp(-\beta \cdot \Delta t)$, where Δt is the time since each prior infusion. Concentrations were evaluated on a dense temporal grid spanning the full 180-minute session for each animal. Each delivery contributed $70 \mu\text{g}/\text{mL} \times 0.05 \text{ mL}$ scaled by that animal's body weight, and each animal's baseline reference curve was taken as the mean of its Days

16–19 trajectories, so that concentrations in later sessions could be expressed as a percentage of that animal’s own stable-baseline peak. The population-average Days 16–19 trajectory was used to identify the characteristic peak in brain fentanyl concentration. We defined the loading phase as the period from session start until the time of this peak (approximately the early part of the session) and the maintenance phase as the remainder of the session. For each animal, we counted how many rewarded lever presses occurred during the loading vs maintenance windows by assigning each press to the appropriate phase based on its timestamp relative to the model-derived peak. These counts were used to summarize how animals distributed their intake between loading and maintenance under baseline conditions and to compare press rates across different shock days. To relate this pharmacokinetic structure to the trial-based paradigm, we also estimated the typical time at which the 15th trial ended (the end of the “Safe” block) by scanning the event logs for trial-end codes and extracting the timestamp of the 15th trial across the Danger sessions. The average of these times, 12.5 ± 1.6 min (mean $\pm$ SD), provided an empirical boundary between the Safe and Danger periods within sessions. For each animal and session, we then computed Safe and Danger phase press counts: presses occurring before this average 15th-trial boundary (early/Safe window) and presses occurring after it (late/shock window).

4.2.3 Change point detection of within-session strategy shifts

Within-session transitions in the self-administration strategy were identified by fitting a Bernoulli changepoint model to each animal’s trial-by-trial active-press train. For each animal-session, the model returned a maximum-likelihood shift trial together with a permutation-based confidence for that detection. Confidence was recorded as a measure of how well resolved each shift was and all animal-sessions were retained. Shift-trial latency was expressed as trials elapsed after Danger onset (trial 16), and retained pressing was expressed as the post-shift press rate as a percentage of the pre-shift rate. Session-level differences in shift latency were tested by within-subject label permutation. Shift latency was not expected to change in one direction across the ladder, since it should rise while pressing is preserved and fall once pressing collapses. A monotone rank correlation is attenuated on such a curve because its rising and falling limbs cancel. Shape was therefore tested directly on the escalation ladder (Relapse excluded) by evaluating the sign of both boundary slopes and the location of the interior turning point. Within-subject deltas were computed by referencing each animal’s shift trial to its own No Shock session and corrected across intensities using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Retained pressing was tested with a single planned contrast between low-intensity sessions (No Shock, 0.1, 0.2, 0.4 mA) and high-intensity sessions (0.8 mA, Relapse) using a sign-flip permutation. Independence of the two axes was assessed by Spearman correlation between each animal’s shock-driven change in shift timing and its change in retained pressing.

For the comparison against vocalization timing, each animal’s shift trial was additionally converted to seconds using that animal’s own trial start timestamps and expressed relative to the start of its first Danger trial, so that the between-animal jitter in when Danger onset occurred

in absolute session time (median 781 s; range of session medians 768–798 s) was removed rather than absorbed into the between-session contrast. Shift latencies in seconds are reported as mean $\pm$ SEM across animals. High-frequency call peaks were obtained from the distribution of >35 kHz call onsets, pooled across animals within each session and binned at 60 s. Local maxima of the raw histogram were identified with `scipy.signal.find_peaks`, retaining maxima whose topographic prominence was at least 15% of that session’s maximum bin count. The histogram was zero-padded at both ends so that first- and last-bin maxima were detectable. Maxima separated by ≤ 200 s were merged to their mean position. Where three or more merged maxima remained, the earliest and latest were discarded, and the latest remaining maximum was taken as that session’s call peak. Peak support was defined as the width at half prominence (`scipy.signal.peak_widths`, `rel_height = 0.5`). The error on each peak is the standard error, across the animals contributing calls within that support ($n = 2\text{--}5$), of each animal’s mean call time referenced to its own Danger onset.

To compare call timing with behavioral strategy timing within animals, each animal-session was assigned a Danger onset equal to the start time of its own trial 16, read from that session’s event record rather than from an assumed trial pacing. High-frequency calls were defined as those exceeding 35 kHz principal frequency, matching the classification used for the session curves. For each animal-session we computed the median high-frequency call time and subtracted that animal’s onset, yielding a signed latency in which negative values precede Danger onset. An empirical median was used rather than a smoothed peak so that the estimate carries no kernel or bandwidth choice and remains defined for animals contributing few calls. Animal-sessions with fewer than five high-frequency calls were excluded, leaving 29 animal-sessions from 8 rats. The strategy changepoint for the same animal-session was converted to seconds by reading the start time of the first post-shift trial, and was referenced to the same onset. Because both quantities were then available for the same animals, they were compared with a two-sided Wilcoxon signed-rank test. The 0.4 mA session contributed a single animal, since vocal suppression at that intensity left no other animal above the call threshold, and it is reported without an error term.

4.2.4 Ultrasonic Vocalization Recording & Analysis

Ultrasonic vocalizations (USVs) were recorded during normal and shocked oral fentanyl self administration sessions using MEMS ultrasonic microphones (M500-384; Pettersson Elektronik). Microphones were inserted through a small port in the chamber wall and housed within a Plexiglas sleeve surrounded by sound-dampening foam to minimize external acoustic contamination. Audio acquisition began at session start and continued for 80 minutes. Signals were digitized at 250 kHz (16-bit) using SASLab Lite (Avisoft Bioacoustics) and saved for offline analysis.

USVs were detected automatically using DeepSqueak v3 (Coffey et al., 2019) with a frequency passband of 18-95 kHz. All automatically detected calls were subsequently reviewed by experimenters blind to behavioral conditions to remove false positives and adjust call boundaries when necessary. Following preprocessing, each call was parameterized using a hybrid feature-extraction approach combining (1) variational autoencoder (VAE) embeddings capturing high-dimensional

spectrotemporal structure (Goffinet et al., 2021), and (2) traditional contour-based acoustic features including duration, start and end frequency, peak frequency, bandwidth, and frequency-modulation statistics (Coffey et al., 2019). Feature sets were concatenated, z-normalized, and subjected to iterative K-means clustering ($K = 1-100$). The optimal cluster count ($K = 9$) was determined using the elbow method applied to within-cluster sum-of-squares. Cluster identities were assigned descriptive labels post-hoc based on characteristic contour shapes.

For affective classification, calls were separated into positive-affect “50-kHz-range” calls and negative-affect “22-kHz-range” calls based on the principal frequency threshold of 35 kHz, following established affective-state literature (Brudzynski, 2013). USV production was aligned to behavioral structure by synchronizing audio timestamps with event logs from the operant system. For cluster-level analyses, per-epoch call counts were normalized to call rate (calls/minute) to account for differences in epoch duration. Cluster distributions and affective call proportions were visualized in low-dimensional acoustic space using UMAP embeddings (McInnes et al., 2020), enabling visualization of call-type segregation and session-dependent shifts in repertoire structure.

4.2.5 Pose Estimation and Kinematic Analysis

Behavioral videos were manually annotated using a custom-built Python graphical interface (Python 3.10, OpenCV 4.8, Tkinter) designed for pose annotation. A custom annotation interface was used to ensure precise control over keypoint definitions, consistent rat-centered cropping in low-light operant chambers, tight alignment with trial structure, and curated sampling of behaviorally relevant frames (e.g., stretch-attend, approach, withholding), producing cleaner training data and more reliable kinematic extraction than would be obtained with generalized pose-tracking frameworks. For each video frame, seven anatomical keypoints were labeled: head center, body center, tail, front right paw, front left paw, left back paw, and right back paw. The annotation tool provided frame-by-frame navigation, drag-and-drop point placement, interpolation between labeled frames, and quality control features including frame-by-frame review and correction capabilities. Annotations were exported in YOLO pose format, with each labeled frame producing: (1) a cropped image containing the rat, and (2) a corresponding text file specifying normalized bounding box coordinates (x_center , y_center , width, height) and keypoint positions (x , y , visibility) for all seven anatomical landmarks. Pose estimation models were trained using the YOLOv8n-pose architecture (Ultralytics YOLOv8) initialized from pretrained COCO-pose weights. The training dataset was split 80/20 into training and validation sets. Models were trained for 5 epochs with early stopping (patience = 3 epochs) using the following hyperparameters: image size = 640x640 pixels, batch size = 16, learning rate = 0.001 (with cosine decay), augmentation enabled (rotation $\pm 10^\circ$, translation $\pm 10\%$, scale $\pm 10\%$). Model performance was evaluated using mean average precision at IoU thresholds of 0.5 (mAP50) and 0.5:0.9 (mAP50-95) for both bounding box and keypoint detection. Final model selection was based on validation set mAP50-95 for keypoints. The best-performing model achieved mAP50 = 0.9942 and mAP50-95 = 0.7657 on the held-out validation set. Trained models were applied to all behavioral videos to generate frame-by-frame predictions of the seven keypoints. For each detection,

the model returned (x, y) coordinates and confidence scores (0-1) for each keypoint. Predictions were retained only if: (1) detection confidence exceeded 0.5 for the bounding box, and (2) all seven keypoint confidence scores exceeded 0.5.

Trial-level movement features were extracted from tracked keypoint trajectories aligned to behavioral event timestamps from the operant system. To test whether gross motor output changed across the imaged sessions, we computed a session-level locomotion measure independent of trial structure. For each session, total locomotion was defined as the cumulative path length of the body-center keypoint via the sum of frame-to-frame Euclidean displacements over all retained frames of the session (pixels; chamber and camera geometry were held constant across the compared sessions, so pixel units are directly comparable across them). Frames failing the confidence criteria above were linearly interpolated before summation so that detection gaps did not inflate displacement.

4.2.6 Surgical Preparation and Chronic Miniscope Imaging

Fourteen days before behavior, rats (n=4 males, n=4 females) were anesthetized with 1-3% isoflurane, placed in a custom robotic stereotaxic instrument (Coffey et al., 2013). The skull leveled to bregma-lambda within ± 0.05 mm. Small craniotomy made above right LHb. Animals received 1000-nL injection of AAV1-EF1 α -jGCaMP8f-WPRE (Addgene #176756) at 100 nL/min via 33-gauge needle (NF33BV-2, WPI). Coordinates: AP -3.65 mm, ML +0.85 mm, DV -5.2 mm for males; females scaled $\times 0.95$. After 10-min diffusion, the needle was withdrawn. GRIN lens (0.6 mm diameter, 6.3 mm length; GRINtech) lowered to DV -4.8 mm (males; scaled for females), secured with jeweler’s screws, cyanoacrylate, and dental acrylic (Parkell C&B Metabond). Animals received meloxicam (2.5 mg/kg s.c.) post-operatively and monitored ≥ 3 days.

Beginning Day 19 and continuing Days 21, 23, 30, calcium imaging conducted using Miniscope V4 (UCLA; www.miniscope.org). Prior to Day 19, animals underwent a dedicated habituation session in which they were connected to the miniscope and allowed to freely explore the chamber, ensuring acclimation to the head-mounted setup and minimizing handling- or equipment-related novelty during imaging days. Miniscope connected via lightweight coaxial cable to data acquisition system (Miniscope DAQ v3.2); video streams saved as .avi with synchronized behavioral timestamps. Custom system for cable protection was used with: (1) strain-relief epoxy at baseplate-cable interface, (2) flexible wire cage surrounding coaxial cable from headstage to commutator. Essential for recordings during high-motion and aversive-stimulation sessions.

4.2.7 Miniscope Processing Suite (MPS) pipeline

All calcium imaging data were processed using the Miniscope Processing Suite (Peden-Asarch et al., 2025), an open-source GUI-based pipeline optimized for long-duration one-photon imaging. Processing parameters were held consistent across animals, with automatic parameter suggestions utilized to optimize performance for individual datasets. Video files were loaded via OpenCV and formatted as xarray datasets with Dask-based chunked arrays for memory-efficient processing. Background subtraction used a morphological tophat filter (window size = 15 pixels) to remove

neuropil contamination. Spatial denoising then applied a median filter (7x7 kernel). Rigid motion correction was performed using FFT-based cross-correlation registration with sub-pixel accuracy along the frame dimension. Erroneous frames containing line-splitting artifacts or motion outliers (>5 standard deviations from median) were automatically detected and removed from further analysis. Interactive ROI cropping reduced the field of view by approximately 60%, excluding non-neural regions based on centered circular crop with manual offset adjustments dependent on each neuronal recording. Spatial-temporal structure was extracted using non-negative double singular value decomposition (NNDSVD) with 100 components and 5 power iterations, applying a sparsity threshold of 0.05. Initial neuronal candidates were identified via watershed segmentation applied to maximum-intensity projections (minimum distance = 10 pixels, relative threshold = 0.1, Gaussian smoothing $\sigma = 1.0$). Spatial components were filtered by size (9-5000 pixels) and merged when spatially proximate (distance threshold = 25 pixels, size-ratio threshold = 5.0, with cross-merge enabled). Temporal signals were extracted via region-averaging with batch processing (10-component batches, 10,000-frame chunks) using memory-efficient modes. Components were further merged based on temporal correlation (threshold = 0.75) and spatial overlap (threshold = 0.3), removing spatially redundant units while preserving distinct neural signals. Per-pixel noise was estimated from background-subtracted residuals using a noise scaling factor of 1.5. Components were validated for size constraints (10-1000 pixels). Temporal signals were then refined with constrained non-negative matrix factorization (CNMF). Neural activity was deconvolved using sparse optimization with autoregressive calcium dynamics (AR-2 model). The algorithm employed noise-weighted L1 penalization (sparse penalty = 0.0229, zero threshold = 6.64×10^{-4}) with normalization enabled. Deconvolution ran for a maximum of 500 iterations per component. Quality control filtered components based on minimum spike activity ($>1 \times 10^{-6}$), temporal variance ($>1 \times 10^{-6}$), and spatial footprint sum ($>1 \times 10^{-6}$). Spatial footprints were refined using dilation-bounded localized LASSO regression. Each component's initial footprint was dilated by 3 pixels (threshold = 0.1) to define a search region. KD-tree algorithms identified neighboring pixels, and multi-penalty LASSO regression (10 penalties ranging from 1×10^{-6} to 0.01) was applied to reconstruct sparse spatial maps. Conservative penalty suggestions (minimum standard deviation = 0.185) were computed from sample frames ($n = 1000$) to balance spatial precision and noise suppression. Growth was constrained to prevent over-expansion. Post-refinement merging removed temporally correlated (threshold = 0.8) and spatially overlapping (threshold = 0.3) components, with smoothing ($\sigma = 1.5$) applied to reduce salt-and-pepper noise. Size filtering (10-5000 pixels) was reapplied. Calcium traces were re-extracted by projecting video data onto refined spatial footprints, followed by a final CNMF deconvolution pass ($p = 2$, sparse penalty = 0.01, zero threshold = 5×10^{-4} , 500 iterations maximum). Final quality filtering retained components meeting minimum size (10 pixels), signal-to-noise ratio ($\text{SNR} > 2.0$), and spatial-temporal correlation (>0.8) thresholds. Processed datasets were exported in Zarr, NumPy, JSON, and pickle formats containing spatial footprints (A), temporal components (C), deconvolved spike estimates (S), and background terms (b, f) with full metadata. All spatial footprints and calcium traces were manually inspected using the MPS Data Explorer interface.

Components were excluded if they exhibited: (1) vascular morphology, (2) spatial footprints outside the field of view boundaries, (3) non-stationary baseline dynamics including sustained step-wise fluorescence increases inconsistent with calcium transients, or (4) anatomical localization outside the lateral habenula based on histological verification.

4.2.8 Stars2Cells: Cross-session registration with Stars2Cells

To track the same LHb neurons across days, spatial footprints from different sessions of the same animal were registered with Stars2Cells (S2C) (Peden-Asarch et al., 2026e). For each within-animal session pair, S2C estimated a global rigid transform (rotation and translation, with scale optionally permitted) between the two sessions’ neurons by RANSAC, fitting the transform on correspondences between local groups of neurons rather than on individual centroids, which is robust to neurons that appear on only one day. Neurons were matched one-to-one across the two sessions by optimal (Hungarian) assignment on their aligned centroid distances, and a pair was counted as the same neuron only when the residual between the assigned centroids fell within 20 pixels, about one cell radius. Across the full segmented populations, S2C matched 373 of 529 neurons from 0.2 to 0.8 mA (70.5%), 222 of 436 from 0.2 mA to Relapse (50.9%), and 298 of 449 from 0.8 mA to Relapse (66.4%). Because our cross-day responsiveness analysis concerned only the reward- and shock-classified neurons, we report the tracking rates over that labeled subset: S2C tracked 153 of 190 neurons from 0.2 to 0.8 mA (80.5%), 109 of 190 from 0.2 mA to Relapse (57.4%), and 138 of 253 from 0.8 mA to Relapse (54.5%). The within-animal session pairs were related by rigid transforms at fixed unit scale, with rotations spanning near-zero to nearly 180°; the large rotations reflect inverted miniscope re-mounts between sessions (5 of 15 adjacent-session registrations exceeded 90°), which S2C recovered automatically by RANSAC and registered across without manual reorientation.

4.2.9 Calcium Signal Quantification and Trial-Aligned Analysis

The strategy-shift trial for each animal-session was taken from the Bernoulli changepoint model (Fig. 4.3), a maximally-selected likelihood-ratio changepoint on a zero-one sequence (Hinkley and Hinkley, 1970) restricted to decreasing splits at or after Danger onset, with significance assessed by permutation of the press train rather than by the asymptotic null. We took the first rewarded trial after that shift as the choice to continue self-administration despite the shock. For each rewarded lever press, we aligned activity to three events, the onset cue at trial start, the rewarded lever press, and the shock, over a window from 5 s before to 2 s after each event at 10 Hz. Each neuron’s trace was z-scored to its own pre-cue baseline over the 5 s before trial start, with the baseline SD floored at 25% of the session median so that quiet neurons were not divided by near-zero noise. Given that animals contributed different neuron counts, traces were rescaled within animal and population means were taken across animals rather than across pooled neurons, so that no single animal dominated a panel. We classified a neuron as responsive to an event if its peak in the first second after the event was positive and exceeded 1 standard deviation. This primary criterion is deliberately permissive relative to the 2–3 SD cutoffs in common use (Zachry et al., 2024; Beacher et al., 2025),

since it serves only to nominate candidate responses that were then gated against a circular-shift null (Kong et al., 2025; Harris, 2020); conventional parametric thresholds assume independence of samples, which calcium transients violate (Harris, 2020). However, since a responsive group can appear in a panel purely because averaging pulled a few noisy traces above threshold, we gated every group against a per-panel null. For each panel we measured the peak separation between the responsive-neuron mean and the whole-panel mean over the first second after the event, and then compared it to the separation obtained after randomly circular-shifting each neuron’s trace. A group was kept only if its separation beat the 99th percentile of 200 such shifts, otherwise it was returned to the non-responsive group. Moreover, to test whether these ensembles were specific to the shocked take, we ran a parallel first-press analysis on the first five rewarded presses of each session, which fall early during loading before any strategy shift. The five presses were averaged within animal and were classified and gated the same way, using the cue and reward events only since these early presses carry no shock ensemble. Additionally, since averaging five presses can smooth noise into an apparent band, we ran the first-press analysis as a split-half, classifying the group on the odd presses and testing its separation on the held-out even presses, so that the presses used to select the group were independent of those used to test it (Kriegeskorte et al., 2009).

For calcium imaging analyses, $n=8$ animals were recorded across Days 19, 21, 23, and 30, giving 32 animal-sessions, of which 2 were unrecoverable owing to line-splitting artifact and were excluded, leaving 30 sessions that yielded 68 ± 20 neurons per session (mean $\pm$ SD; median 68, range: 33-112) and 2,027 neuron-sessions in total. Transients were detected using conservative thresholds (baseline + 3.5 SD, prominence 2.5 SD, minimum distance 10 frames / 1.0 s at 10 Hz). Within a session a neuron keeps one identity, so the cue, reward, and shock labels for a given neuron already refer to the same neuron and no registration is needed. Thus, for each session we took every neuron with a valid label in both epochs and asked what fraction of the neurons responsive to the first epoch were also responsive to the second. The step was cue to rewarded lever press at 0.2 mA and rewarded lever press to shock at 0.8 mA and at Relapse. Overlap was compared against the marginal responsiveness rate of the second epoch, that is, the fraction of all neurons validly labeled in both epochs that were responsive to the second, which is the overlap expected if the two labels were independent. Two-sided significance came from a permutation that shuffled the conditioning label within animal (10,000 draws, preserving each animal’s marginal rate so that the test does not pseudoreplicate across rats), corroborated by a two-sided Fisher’s exact test on the pooled 2×2 table; fractions carry Wilson 95% confidence intervals. We used the S2C outputs for tracking cross-day stability. For each within-animal session pair we linked neurons, kept only neurons that were both tracked and labeled in both sessions, and looked into what fraction of the neurons responsive in the earlier session stayed responsive in the later one. Neurons responsive in the later session that tracked to no earlier neuron or that tracked to an earlier neuron whose label failed QC, were dropped rather than scored as turnover. We focused on the responsive neurons to the rewarded lever press from 0.2 to 0.8 mA and from 0.8 mA to Relapse, and to the shock from 0.8 mA to Relapse. Cross-day overlap was tested identically, against the marginal responsiveness rate of the later session computed

over all neurons tracked and validly labeled in both.

Peak separation was defined as the maximum, over the first second after the event, of the absolute difference between the animal-weighted mean trace of the responsive neurons and the animal-weighted mean trace of the whole panel, in units of the pre-cue baseline SD. Its null was generated by independently circular-shifting each neuron’s trace by a random offset, re-deriving the responsive set on the shifted data, and recomputing the same quantity (200 shifts per panel), so that the selection step is applied identically to observed and surrogate data; a group was retained only if its observed separation exceeded the 99th percentile of that null. For the split-half first-press analysis the responsive set was instead learned on the odd presses and held fixed while both the observed separation and its shift null were computed on the held-out even presses. Response magnitude was the signed peak z-score of each neuron in the first second after its event, and changes in magnitude between epochs or between sessions were tested on the paired per-neuron values with a two-sided Wilcoxon signed-rank test.

4.2.10 Population decoding of task events and adjacent-trial choice

Beyond the responsivity classification above, we asked directly what the LHb population represents, using regularized decoders scored against explicit chance. All decoders were L2-penalized logistic regressions on z-scored population activity with balanced class weights. An L2-penalized logistic regression was the right fit because a session yields tens of trials but hundreds of neurons, and the penalty keeps the weights small and spread across neurons that covary rather than letting one cell carry the decision. A linear decoder can only use signal that survives a weighted sum of firing rates, which is also the operation a downstream target performs most readily. A more flexible classifier could use patterns that a weighted sum cannot express, and would score at least as well. The AUCs we report are therefore a floor on the information present in the LHb population rather than a ceiling. The penalty strength was fixed at $C = 1.0$ for every decoder in the study, so no hyperparameter was tuned on these data and no model selection step could leak into the reported scores. Standardization was fit inside each training fold and applied to the held-out fold, so the scaler never saw the test trials. Class weights were balanced because presses and withholds are unequal in number and their ratio varies by session. Each decoder was scored by out-of-fold ROC-AUC under stratified cross-validation with up to five folds. The fold count was set to the size of the minority class when that was fewer than five, so every fold contained both classes. Held-out predictions were pooled across folds before scoring rather than averaging fold-wise AUCs, which is unstable at these trial counts. ROC-AUC was the primary metric because it is threshold-free and insensitive to class prevalence, and average precision and balanced accuracy were retained alongside it. Each decoder was attempted only when at least four trials of both classes were available, and all randomization was seeded.

For each imaging session and each event (trial onset, onset cue, rewarded lever press, shock), we decoded the population activity vector at every 0.1 s lag from -2 to $+2$ s around the event against a time-matched pre-event baseline. The baseline was a single frame 4 s before the same event, which

holds trial context constant. Event decoders were run on two neuron sets in parallel, all segmented neurons in a session and the event-responsive ensembles defined above. Decoders were trained per animal-session and the resulting ROC-AUC curves were averaged across the eight animals. Training within animal is required rather than preferred, since neuron identity is not shared across animals and the feature space is therefore animal-specific, which makes the animal the unit of inference. We report the value at the event ($t = 0$) as the unbiased estimate, since the peak across lags is upward-biased by selection.

To ask whether within-trial activity predicts the choice on the neighboring trial, we focused on the isolated-press (miss, hit, miss) motif and defined three per-animal decoders on all segmented neurons. Features were the mean population activity in the 1 s after a source event. The first decoder, *before-miss*→*press*, read a withhold trial’s onset or cue and predicted whether the next trial was a press (miss→hit vs miss→miss). The second, *after-miss*←*press*, read a withhold trial’s onset or cue and decoded whether a press had occurred on the immediately preceding trial (hit→miss vs miss→miss). The third, *in-hit*→*stop*, read a press trial’s onset, cue, rewarded press, or shock and predicted whether the next trial was a withhold (hit→miss vs hit→hit). Each decoder was run over all trials, separately within the Safe and Danger phases, and separately within the pre- and post-shift epochs. Chance was established empirically rather than assumed, by permuting the trial labels and refitting the full cross-validation pipeline on each permutation, which preserves class balance and the covariance among neurons while destroying only the label mapping. Group-level significance was assessed across animals by a one-sided Wilcoxon signed-rank test of the per-animal AUCs against chance (0.5). Stouffer’s method combined the per-animal label-shuffle nulls (100 permutations each) as a complementary test.

Simulated photometry.

To ask what a bulk signal would have reported from the same recordings, we collapsed every segmented neuron in each field of view into a single trace, averaging across neurons and across all rewarded trials of the session, following prior work that approximates a fiber signal by summing cellular-resolution recordings with uniform weights (Legaria et al., 2022; Yun and Parker, 2025). Because that approximation retains only segmented somata, it is the somatic limit of a fiber rather than a full field average, and a fiber additionally sums neuropil that carries much of the bulk signal (Legaria et al., 2022). Traces were z-scored to the same pre-cue baseline used for the ensemble analysis, aligned from -2 to $+2$ s around each event to match the decoders’ lag axis, and referenced to the 1 s preceding the event. The bulk statistic was the mean over the first second after the event, in baseline SD, and was tested across the eight animals by a two-sided Wilcoxon signed-rank test against no change.

4.2.11 Histology

Upon experiment completion, animals were anesthetized (phenytoin/pentobarbital, 100 mg/kg i.p.), transcardially perfused with ice-cold PBS then 4% paraformaldehyde. Brains were post-fixed (24 h, 4°C, 4% PFA), processed using iDISCO+ protocol (Renier et al., 2014). Serial methanol

dehydration (20%, 40%, 60%, 80%, 100%, 100%; 1 h each, RT), permeabilization (5% H₂O₂ in 100% methanol, overnight, RT), rehydration (100%, 80%, 60%, 40%, 20% methanol in PBS-T; 1 h each, RT). Samples incubated in 20% DMSO/0.3M glycine/0.2% Triton X-100/PBS (2 days, 37°C), blocked (10% DMSO/6% donkey serum/0.2% Triton X-100/PBS, 2-3 days, 37°C). Primary antibody: chicken anti-GFP (Aves Labs #GFP-1020; 1:750 in 5% DMSO/3% donkey serum/0.2% Tween-20/PBS with 10 μ g/mL heparin, 7 days, 37°C), washed 3x24 h in 0.2% Tween-20/PBS with 10 μ g/mL heparin (37°C). Secondary antibody: goat anti-chicken Alexa Fluor 647 (ThermoFisher #A-21449; 1:500, 7 days, 37°C), washed 3x24 h. Final clearing: methanol dehydration (20%, 40%, 60%, 80%, 100%, 100%; 1 h each), equilibration (67% dichloromethane/33% methanol, 3 h, RT), DCM washes (2x15 min), immersion in dibenzyl ether (Sigma #108014, 2-3 days, RT).

Whole-brain clearing was performed using the SmartBatch+ system (LifeCanvas Technologies). Briefly, fixed brain tissue blocks (~ 1 cm³, centered on the lateral habenula) were processed using the SHIELD protocol (stabilization under harsh conditions via intramolecular epoxide linkages) to preserve tissue architecture and fluorescence (Park et al., 2018). Samples were then cleared for 6 days using standard SmartBatch+ reagents and protocols (LifeCanvas Technologies, n.d.), followed by optical refractive index matching with EasyIndex (LifeCanvas Technologies; #EI-500). Cleared brains were imaged on an Ultramicroscope II large-format upright light-sheet microscope (Mitenyi Biotec), and volumetric reconstructions were generated in Fiji (ImageJ). The GRIN lens tract in each brain included in the analysis was manually localized in three dimensions via visual inspection and registered to the rat brain atlas (Paxinos and Watson, 2014).

4.2.12 Statistical Analyses

All statistical analyses were performed in Python 3.10 with `scipy.stats` and `numpy`. Data are presented as mean $\pm$ SEM, significance was set at $\alpha = 0.05$, tests were two-sided unless stated otherwise, and exact p-values are reported unless $p < 0.001$. Sex differences over the Days 16-19 stable baseline were tested with independent samples t-tests comparing males ($n=16$) and females ($n=11$) on total rewarded lever presses, head entries, latency, active lever presses, inactive lever presses, and weight-normalized intake, with Bonferroni adjustment for six comparisons ($\alpha = 0.0083$). Danger-phase analyses used the 27 animals that ran the delayed shock paradigm. Friedman tests were applied to Safe-trial responding (trials 1-15) during shock escalation (Days 19-22), to Danger-trial responding (post-trial-15) across all shock sessions (Days 19-23 and the Day 30 Relapse), and to press rates during the loading phase (0 to 12.5 min) and the maintenance phase (12.5 to 180 min). The Danger-trial test was followed by post-hoc pairwise Wilcoxon signed-rank tests, and each shock session was also compared with long-term baseline (Days 1-19 average) by the same test. Inference on the changepoint outputs rested on permutation rather than on parametric models. The maximally-selected likelihood ratio has no asymptotic χ^2 reference because the shift trial is unidentified under the single-rate null, so each detection was calibrated against 2000 permutations of that animal's own press train. Shift lag is an overdispersed count, with a variance-to-mean ratio of 3.7 to 58.5 across sessions, and a Poisson generalized estimating equation with cluster-robust

standard errors returned a 26.7% false-positive rate at a nominal 5% level under a matched null, reflecting the downward bias of the sandwich estimator with few clusters. Sign-flip permutation was used for the paired session contrasts because an exact zero contributes zero regardless of sign and is retained, rather than discarded as a Wilcoxon signed-rank test would do. The non-monotonicity test on the escalation ladder was the Sasabuchi and Lind-Mehlum procedure, whose joint p-value is the larger of the two one-sided boundary tests. For continuity with earlier reporting we also fit a linear mixed-effects model of shift lag on session with a subject random intercept, though its random-effects covariance was singular, so its Holm-corrected pairwise contrasts are treated as a separate family and were not relied upon. High-frequency calls and low-frequency calls were compared across sessions with Friedman tests ($n=8$), and total calls per session were compared with Wilcoxon signed-rank tests.

4.3 Results

4.3.1 Oral fentanyl self-administration produces stable baseline intake independent of sex

We trained male ($n=19$) and female ($n=14$) Long-Evans rats to self-administer liquid fentanyl (70 $\mu\text{g}/\text{mL}$, 0.05 mL/delivery) using a saccharin-fade procedure for five days with 19 days of acquisition access (Fig. 4.1a–d).

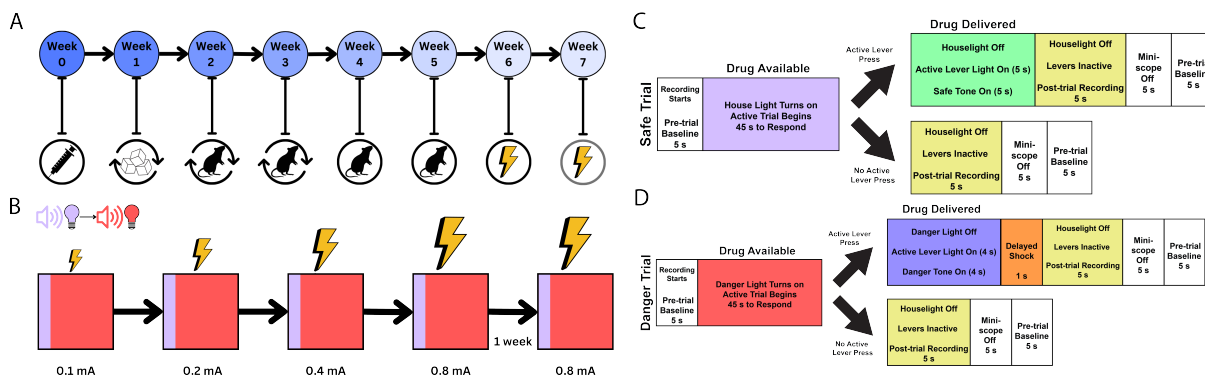


Figure 4.1: A delayed-shock paradigm for fentanyl self-administration. (A) Week-level experimental timeline: viral injection (Week 0), saccharin training (Week 1), saccharin fade and acquisition (Weeks 2–5), daily shock escalation (Week 6), and Relapse (Week 7). (B) Trial structure on Safe trials: an active lever press triggers fentanyl delivery with no programmed consequence. (C) Trial structure on Danger trials: an active lever press triggers fentanyl delivery followed 4 s later by a 1 s shock. (D) Within-session structure: 15 Safe loading trials (trials 1–15) precede 165 Danger trials (trials 16–180) at an intensity that escalated across days (0.1, 0.2, 0.4, 0.8 mA), followed one week later by a Relapse at 0.8 mA.

Training progressed through four phases: Days 1–5, continuous reinforcement, Days 6–10, continuous reinforcement, Days 11–15, limited hold, and Days 16–19, limited hold. Both sexes rapidly acquired self-administration, with total rewarded lever presses escalating across training (Fig. 4.2a) alongside converging head entries to the reward port (Fig. 4.2d) and active lever pressing (Fig.

4.2b). Inactive lever pressing (Fig. 4.2b) and latency to press (Fig. 4.2d) remained low throughout training, indicating discrimination between active and inactive levers and successful acquisition of the behavioral paradigm. During the stable baseline period (Days 16-19), rats averaged 25.0 ± 1.7 rewarded lever presses, 98.0 ± 9.1 head entries, and 28.4 ± 6.7 s latency from lever press to head entry, demonstrating consistent operant performance. Comparing males and females during this stable baseline revealed no significant sex differences in total rewarded lever presses ($t(25)=0.44$, $p=0.665$), head entries ($t(25)=-0.40$, $p=0.691$), latency ($t(25)=-1.00$, $p=0.328$), active lever presses ($t(25)=-0.07$, $p=0.948$), or inactive lever presses ($t(25)=-0.56$, $p=0.580$) (Fig. 4.2a,b,d). The only sex difference that survived Bonferroni correction ($\alpha=0.0083$) emerged in weight-normalized intake ($t(25)=-3.68$, $p=0.001$), with females consuming 0.30 ± 0.03 mg/kg versus males 0.18 ± 0.02 mg/kg during Days 16-19.

4.3.2 Rats suppress fentanyl consumption abruptly at high shock intensity

We introduced a delayed shock beginning Day 20. Each session started with 15 “Safe” trials allowing drug loading, followed by trials where lever presses resulted in drug delivery followed by a 1 s shock 4 s after an active lever press (Fig. 4.1b–c). Shock intensity escalated daily: 0.1 mA (Day 20), 0.2 mA (Day 21), 0.4 mA (Day 22), 0.8 mA (Day 23), with Day 30 serving as Relapse (0.8 mA repeated) (Fig. 4.1d). Pharmacokinetic modeling placed the peak in brain fentanyl concentration at 12.2 min ($123.2 \mu\text{g/L}\cdot\text{kg}$), coincident with the opening of the Danger phase at 12.5 min. At Danger period onset, the estimated concentration remained at 58 to 88% of each animal’s own Days 16–19 baseline peak across the Danger sessions from 0.1 to 0.8 mA and at Relapse, and remained at or above half of that peak for at least 8 min thereafter.

Danger-trial responding (post-trial-15) showed significant variation across sessions (Friedman test: $\chi^2(5) = 59.89$, $p < 0.001$; Fig. 4.2f). Animals showed a trend toward escalation of post-15-trial responding on Day 20 (0.1 mA; 20.5 ± 3.3 rewarded lever presses) compared to Day 19 baseline (15.1 ± 1.6 rewarded lever presses; Wilcoxon signed-rank: $W = 100$, $p = 0.097$; Fig. 4.2f & Fig. A.1a), with a comparable trend against the long-term baseline average across Days 11-19 (15.3 ± 1.3 rewarded lever presses; $W = 92$, $p = 0.058$).

Contrary to a smoothly graded dose-response, Danger-trial responding was unchanged across the moderate intensities before collapsing at high intensity (Friedman test across Days 19-22: $\chi^2(3) = 1.94$, $p = 0.59$). Day 21 (0.2 mA) did not differ from the long-term baseline (17.2 ± 2.5 rewarded lever presses; $W = 139$, $p = 0.54$; Fig. 4.2f & Fig. A.1b), nor from the preceding 0.1 mA session ($W = 149$, $p = 0.75$). Day 22 (0.4 mA) produced comparable responding (18.8 ± 2.6 rewarded lever presses; $W = 112$, $p = 0.17$ vs. baseline; $W = 146$, $p = 0.66$ vs. 0.2 mA). The 0.8 mA (Day 23) session was the first intensity step that produced a significant drop in rewarded lever presses relative to the preceding session (1.3 ± 0.5 rewarded lever presses; $W = 16$, $p < 0.001$ vs. 0.4 mA; $W = 0$, $p < 0.001$ vs. baseline; Fig. 4.2f & Fig. A.1b). Moreover, suppression deepened further at Relapse, demonstrating a stable conflict-gated decision policy rather than a transient, session-specific effect (0.8 mA repeated; 3.3 ± 0.5 rewarded lever presses; $W = 34$, $p = 0.001$ vs. Day 23; $W = 1$, $p <$

0.001 vs. baseline; Fig. 4.2f & Fig. A.1e).

Safe-trial responding (trials 1-15) remained stable across the escalation ladder (Days 19-22; Friedman test: $\chi^2(3) = 5.35$, $p = 0.15$; Fig. 4.2f) and across all six sessions, including 0.8 mA and Relapse (6.6 ± 0.6 presses at 0.4 mA versus 5.4 ± 0.7 at 0.8 mA and 6.6 ± 0.6 at Relapse; Friedman across all six sessions: $\chi^2(5) = 8.32$, $p = 0.14$). Trial-by-trial rasters revealed substantial inter-animal variability (Fig. 4.2e), with some rats showing rapid all-or-none transitions and others showing gradual response decrement, indicating individual differences in shock sensitivity. In Fig. 4.2f, asterisks mark within-session differences between Safe- and Danger-trial responding, which was higher on Danger trials at 0.1, 0.2, and 0.4 mA, did not differ at 0.8 mA, and was lower on Danger trials at Relapse (** $p < 0.01$).

We separately compared rewarded lever press counts against the Days 1-19 baseline within the loading (0 to 12.5 min) and maintenance (12.5 to 180 min) periods of each shock session (Mann-Whitney U; Fig. A.1f). At lower intensities, maintenance-period pressing was statistically indistinguishable from baseline (22.8 ± 1.1 presses) at 0.1 mA (27.2 ± 5.5 presses, $p = 0.531$), 0.2 mA (20.0 ± 3.3 presses, $p = 0.762$), and 0.4 mA (18.4 ± 2.3 presses, $p = 0.821$), then fell sharply at 0.8 mA (5.2 ± 0.8 presses, $p < 0.001$) and at Relapse (3.2 ± 0.4 presses, $p < 0.001$). Loading-period pressing was more variable, falling below baseline (9.2 ± 0.3 presses) at 0.2 mA (6.2 ± 1.0 presses, $p = 0.012$) and 0.8 mA (6.1 ± 0.8 presses, $p = 0.020$) but not at 0.1 mA (7.6 ± 0.6 presses, $p = 0.680$), 0.4 mA (7.1 ± 0.7 presses, $p = 0.249$), or Relapse (6.9 ± 0.6 presses, $p = 0.209$). These loading-period differences were not ordered by intensity and did not survive Bonferroni correction across the ten comparisons ($\alpha = 0.005$), and thus, we interpret them as reflecting session-to-session variability in loading rather than a graded effect of shock on Safe-period behavior.

Contrary to motor impairment predictions, overall movement during the Danger phase showed modest increases rather than decreases. Safe-trial movement trended upward across days (Day 19: 1488 ± 108 cm/trial; Day 23: 2250 ± 461 cm/trial; $F(3,56) = 1.76$, $p = 0.17$), while Danger-trial movement increased significantly (Day 19: 1985 ± 24 cm/trial; Day 23: 2243 ± 29 cm/trial; $F(3,5481) = 18.03$, $p < 0.001$). Animals therefore moved more while withholding lever pressing, which is consistent with an active defensive or arousal state such as risk assessment or agitation rather than with motor incapacitation (Bailey and Crawley, 2009; Holly et al., 2016; Chu et al., 2024). Latency from lever to port, on the trials where presses occurred, did not increase on high-intensity days (Day 23: 19.3 ± 1.2 s vs. Day 19: 18.7 ± 1.0 s, $p = 0.61$), and animals readily pressed the lever during Safe trials on all days, so withholding during Danger trials reflected active decision-making rather than an inability to respond.

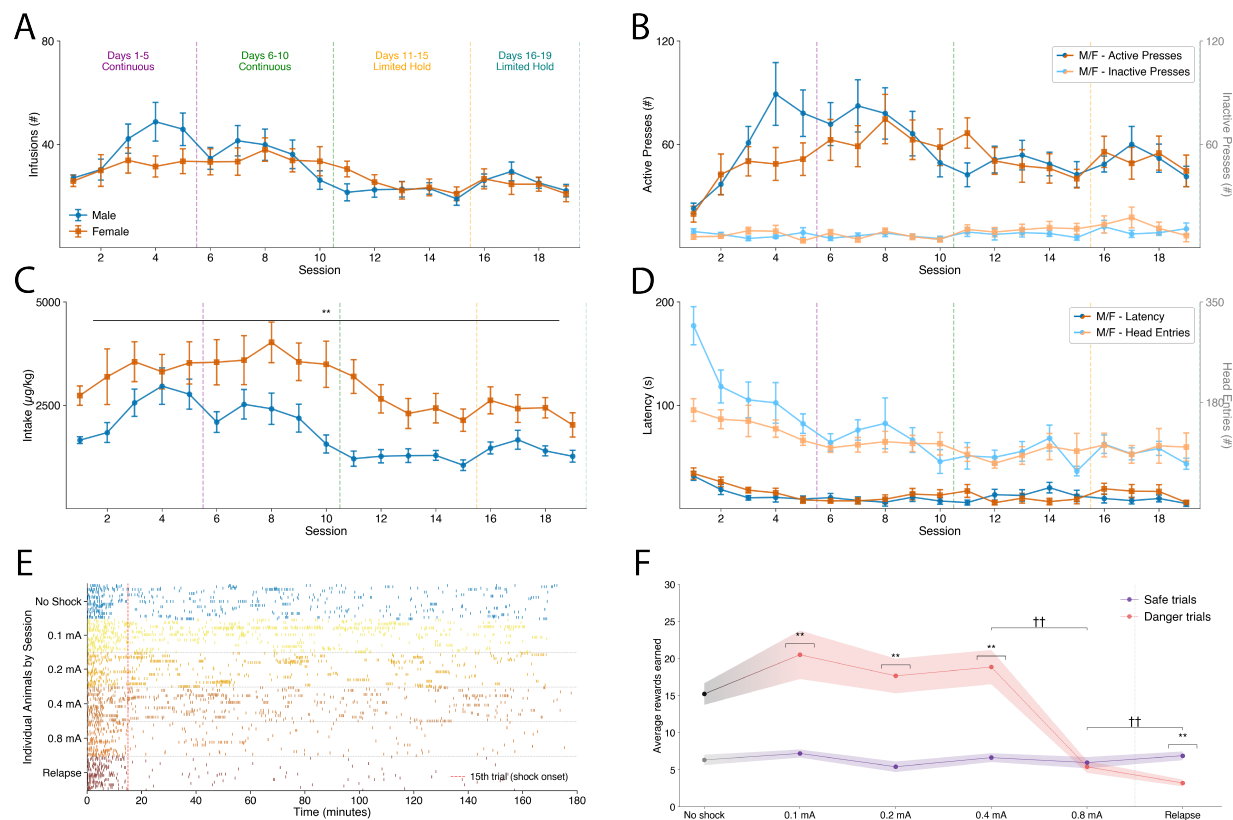


Figure 4.2: Rats suppress fentanyl consumption abruptly at high shock intensity. (A) Total rewarded lever presses per session across the 19 days of training, split by sex. (B) Active (left axis) and inactive (right axis) lever presses per session across training, split by sex. (C) Weight-normalized fentanyl intake ($\mu\text{g}/\text{kg}$) per session across training, split by sex. Cumulative intake, quantified as the area under each animal's intake-by-session curve, was 1.68-fold higher in females ($54204.7 \pm 4874.2 \mu\text{g}/\text{kg}$) than in males ($32240.3 \pm 3172.1 \mu\text{g}/\text{kg}$; Mann-Whitney U, $p = 0.001$). (D) Latency to head entry (left axis) and head entries to the reward port (right axis) per session across training, split by sex. (E) Rasters of rewarded lever presses, blocked by session (No Shock, 0.1, 0.2, 0.4, 0.8 mA, and Relapse). Each row is one animal and each tick is a rewarded lever press; x-axis, time (min). (F) Average rewards earned per session on Safe trials (trials 1–15; grey at No Shock, purple at all shocked sessions) and Danger trials (trials 16–180; black at No Shock, red at all shocked sessions), across the escalation ladder and Relapse. Asterisks, within-session Safe versus Danger ($** p < 0.01$); daggers, between-session Danger-trial change ($\dagger\dagger p < 0.01$). Data are mean $\pm$ SEM across animals; $n=33$ animals.

4.3.3 Shock shifts the timing of a within-session change in strategy

The decrease of rewarded lever presses did not unfold uniformly across a session in response to shock as demonstrated by individual animals continuing to press for a variable number of Danger trials before abruptly reducing their press rate. A Bernoulli changepoint model fit to each animal's trial-by-trial press train resolved a shift at the 5% level in the large majority of animal-sessions in every condition, from 81% under No Shock to 96% at Relapse (Fig. 4.3a–f). Averaged across all animals, shift-trial latency traced an inverted-U along the escalation ladder, rising from 9.2 ± 3.1 trials after Danger onset under No Shock to a peak of 49.2 ± 9.9 trials at 0.2 mA, falling back to 30.2 ± 8.2 trials at 0.4 mA and 12.2 ± 4.9 trials at 0.8 mA, then collapsing to 2.2 ± 0.6 trials at Relapse (within-subject label permutation, 154 animal-sessions, $p = 0.0002$; Fig. 4.3g). Excluding the Relapse, both boundary slopes were individually signed, positive at the bottom of the ladder ($+34.32$, $p < 0.0001$), negative at its top (-33.68 , $p < 0.0001$), and an interior turning point falling between 0.2 and 0.4 mA (joint $p < 0.0001$, 128 animal-sessions). The same non-monotonic pattern held when each animal's shift trial was referenced to its own No Shock session. Relative to their own No Shock baseline, animals switched a median of 2 trials later at 0.1 mA ($p_{\text{FDR}} = 0.027$), 11 trials later at 0.2 mA ($p_{\text{FDR}} = 0.003$), and 4 trials later at 0.4 mA ($p_{\text{FDR}} = 0.030$), whereas the delay was abolished at 0.8 mA (median 0 trials, $p_{\text{FDR}} = 0.65$) and reversed at Relapse (mean 7.0 trials earlier, median 0 trials, $p_{\text{FDR}} = 0.028$).

Additionally, once an animal shifted, the press rate it retained afterward stayed near a fifth of its pre-shift rate across the whole lower half of the ladder, at $18.5 \pm 2.0\%$ under No Shock, $17.4 \pm 3.6\%$ at 0.1 mA, $12.0 \pm 2.4\%$ at 0.2 mA and $14.9 \pm 2.3\%$ at 0.4 mA, dropping sharply to $5.9 \pm 1.0\%$ at 0.8 mA and $3.3 \pm 0.5\%$ at Relapse. The contrast between low- and high-intensity sessions supported this with retained pressing averaged 14.6% across low intensities against 4.6% across high, a drop of 9.9 percentage points that held in 22 of 26 animals (sign-flip permutation, $p < 0.0001$).

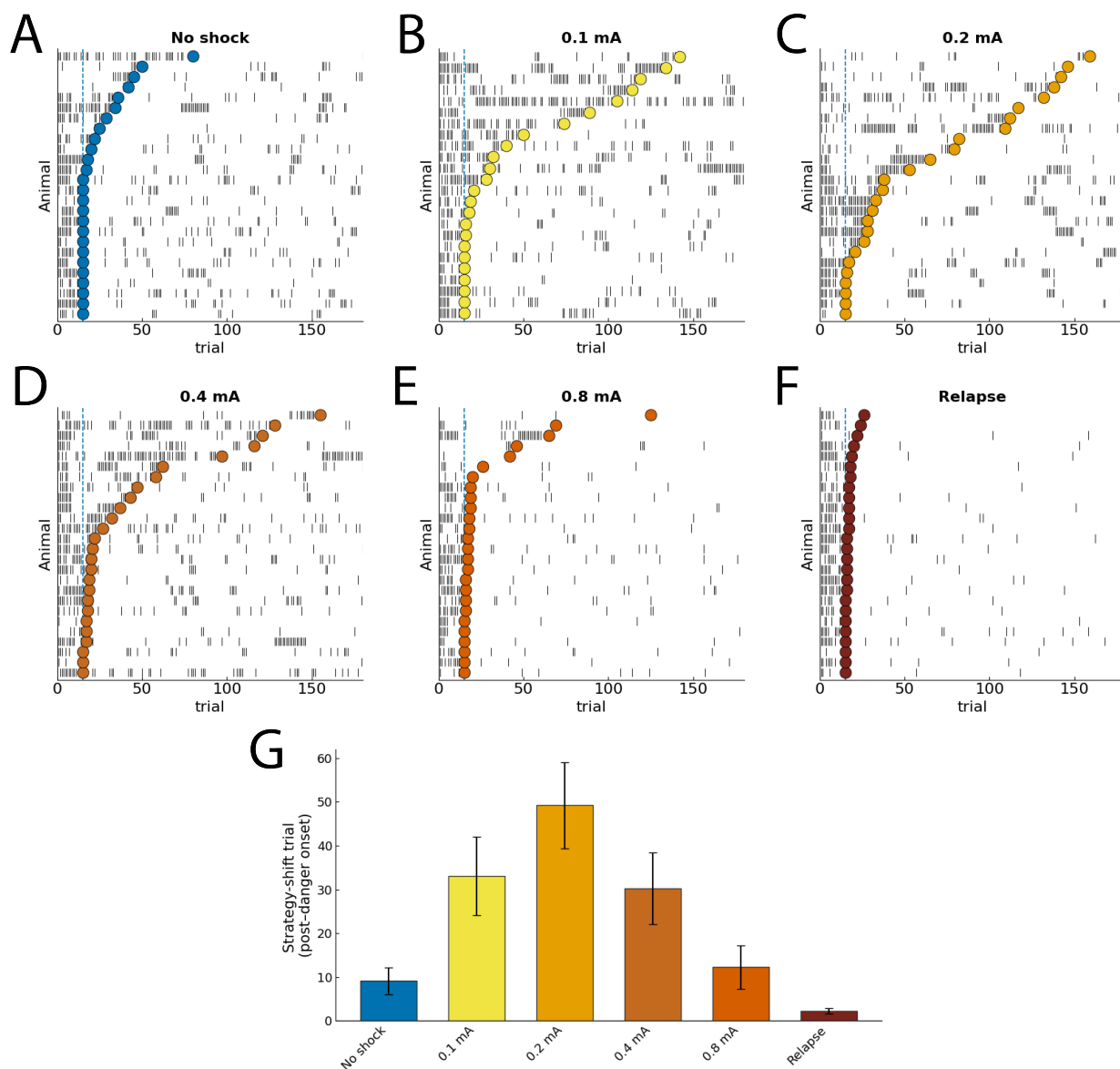


Figure 4.3: Shock shifts the timing of a within-session change in strategy. (A) Trial-by-trial active-press raster for the No Shock session. Each row is one animal, sorted by strategy-shift trial; black ticks are active presses. Filled circles mark that animal’s Bernoulli changepoint-detected strategy-shift trial, shown for every animal without a significance filter. Blue dashed line, Danger onset (trial 16). (B) The same for 0.1 mA. (C) The same for 0.2 mA. (D) The same for 0.4 mA. (E) The same for 0.8 mA. (F) The same for the Relapse. (G) Mean strategy-shift trial relative to Danger onset, one point per session. Tested by within-subject label permutation. Data are mean $\pm$ SEM across animals; $n=27$ animals, 154 animal-sessions.

4.3.4 Ultrasonic vocalizations reveal complex affective dynamics during conflict-ing decision-making

To assess affective state during shock-based decision-making, we analyzed ultrasonic vocalizations (USVs) using DeepSqueak. Unsupervised clustering identified 9 distinct call types (Fig. A.2a). The most prevalent call type shifted across sessions in that baseline sessions favored Trill calls (27.1%), while Days 23 (0.8 mA) and 30 (Relapse) were dominated by Long 22-kHz calls (32.6% and 29.5%, respectively), a sustained low-frequency call type (1051 ± 24 ms duration, $n=830$ calls) associated with negative affective states (Fig. A.2b).

Baseline sessions (Day 19, Safe) showed high-frequency calls (>35 kHz, associated with positive affect) dominating early periods (59.1 ± 25.7 calls per subject, 0-15 minutes) with minimal low-frequency calls (<35 kHz, associated with negative affect; 0.1 calls per subject, 30-60 minutes; Fig. 4.4a). Day 20 (0.1 mA) showed reduced overall calling (45.9 ± 33.2 vs. 80.8 ± 32.4 calls per subject baseline), but maintained the temporal pattern of early high-frequency calls along with a later stage peak (Fig. 4.4b). Day 21 (0.2 mA) preserved high-frequency calling patterns similar to baseline (54.8 ± 25.1 calls, 0-15 minutes) with minimal low-frequency calls throughout (Fig. 4.4c). Day 22 (0.4 mA) produced dramatic suppression of all USV production (3.8 ± 2.9 vs. 80.8 ± 32.4 calls per subject baseline, $p=0.0078$, Wilcoxon signed-rank test; Fig. 4.4d), with 8/8 animals showing reduced calling. This near-complete vocal suppression occurred at an intensity where Danger responding was still at baseline.

Day 23 (0.8 mA) showed recovery of high-frequency calls during early Safe trials (62.6 ± 44.3 calls, 0-15 minutes), followed by emergence of low-frequency calls in late periods (13.1 calls per subject, 30-60 minutes vs. 0.1 baseline; Fig. 4.4e). Long 22-kHz calls dominated this late window (29.0 ± 29.0 calls per subject, 25-40 minutes). Total calling increased significantly from Day 22 (3.8 ± 2.9 to 140.6 ± 100.7 calls per subject, $p=0.0312$). Day 30 (Relapse) displayed a bimodal low-frequency pattern (Fig. 4.4f): an early peak during initial trials (17.2 calls per subject, 0-15 minutes) and a larger late peak during sustained abstinence (37.4 calls per subject, 30-60 minutes; 2.17x ratio). High-frequency calls showed modest variation across sessions (Friedman $\chi^2(5)=14.03$, $p=0.015$), while low-frequency calls showed substantial individual variability but no significant omnibus effect (Friedman $\chi^2(5)=10.25$, $p=0.069$).

High-frequency (>35 kHz) calls peaked within the first 18 min of every session (absolute peak times: 0.2 mA, 30 s; 0.4 mA, 270 s; 0.1 mA, 510 s; Relapse, 690 s; No Shock, 930 s; 0.8 mA, 1110 s), corresponding to the drug loading period rather than to the onset of shock (mean onset across sessions, 781 s). Relative to each session's Danger onset, peaks fell before onset in four of six sessions (-760 s to -80 s) and after onset in two (No Shock, $+160 \pm 12$ s; 0.8 mA, $+319 \pm 18$ s; mean $\pm$ SEM across animals contributing calls to the peak window). By contrast, the behavioral strategy changepoints followed Danger onset in every session (480 ± 162 s at No Shock to 2692 ± 571 s at 0.2 mA) and were separated from call peak timing in every session. High-frequency vocalization therefore tracked the appetitive loading phase of the session and was dissociated from the shock-induced shift in response strategy. Moreover, since session-level peaks group calls across animals, we repeated

this comparison within animals. For every animal-session with at least five high-frequency calls we took the median call time and that animal's own strategy changepoint, each referenced to that animal's own Danger onset. Across 29 animal-sessions from 8 rats, high-frequency calls preceded Danger onset by 149 ± 170 s while the strategy changepoint followed it by 1009 ± 303 s (mean $\pm$ SEM; Wilcoxon signed-rank, $W = 67$, $p = 0.0007$; Fig. 4.4h). Calls fell before Danger period onset in 22 of 29 animal-sessions and changepoints fell after the Danger period onset in 23 of 29. The changepoint in this vocalizing subset was within 206 s of the full-cohort value (1216 ± 169 s, $n = 154$ animal-sessions).

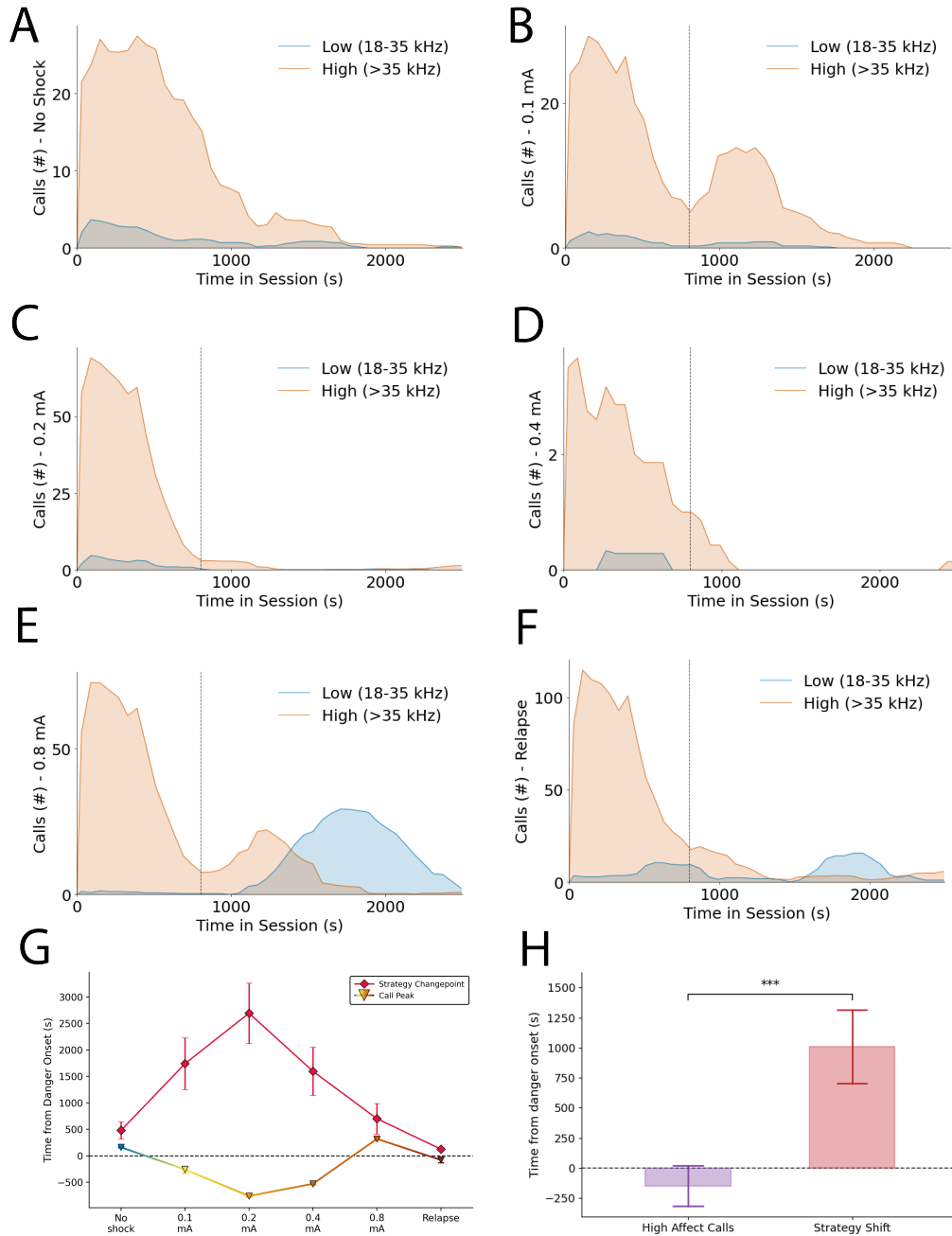


Figure 4.4: Ultrasonic vocalizations reveal complex affective dynamics during conflicting decision-making. (A) Temporal distribution of low-frequency (18–35 kHz, blue) and high-frequency (>35 kHz, red) calls across the No Shock session (Day 19). Shading fills each trace to zero and is not an error band. (B) The same for Day 20 (0.1 mA). Black dashed line, approximate Danger onset (800 s). (C) The same for Day 21 (0.2 mA). (D) The same for Day 22 (0.4 mA). (E) The same for Day 23 (0.8 mA). (F) The same for Day 30 (Relapse). Total calls per session were compared across sessions by Wilcoxon signed-rank test. $n=8$ animals. (G) Timing of the high-frequency (>35 kHz) call peak (triangles, color-graded by session) and of the behavioral strategy changepoint (crimson diamonds), each expressed relative to that animal's own Danger onset (dashed line, $t = 0$; positive values follow onset). Change point values are mean $\pm$ SEM across animals ($n=27$ animals). (H) Median latency of high-frequency calls (purple) and of the behavioral strategy changepoint (red), each relative to that animal's own Danger onset (dashed line, $t = 0$). Compared by two-sided Wilcoxon signed-rank test on the same animal-sessions. Bars are mean $\pm$ SEM across 29 animal-sessions from 8 animals contributing at least five high-frequency calls.

4.3.5 Chronic miniscope imaging and cross-session registration of LHb neurons

We performed chronic miniscope imaging of LHb neurons expressing jGCaMP8f across Days 19, 21, 23, and 30. Raw single-photon miniscope videos were processed through MPS (Peden-Asarch et al., 2025), yielding 68 ± 20 segmented neurons per session (mean $\pm$ SD; median 68, range: 33–112; $n = 30$ sessions across 8 animals), with per-session yield stable across the imaging window (Spearman $\rho = -0.28$, $p = 0.14$) and no session-level drift in a two-way model (count $\sim$ animal + session; session $F(4,18) = 0.65$, $p = 0.64$), the residual between-animal variation reflecting prep-level differences in lens placement and expression ($\text{ICC}(1) = 0.36$; Fig. A.3a–b). This between-animal component accounted for only a minority of the total variance, with within-animal day-to-day variation exceeding it in absolute terms ($\text{SD}_{\text{between}} \approx 12$ versus $\text{SD}_{\text{within}} \approx 16$ neurons), indicating that most session-to-session variation reflected day-to-day segmentation noise rather than stable animal identity. MPS processed the multi-hour recordings (170 ± 5 minutes per session; $101,721 \pm 2,705$ frames at 10 Hz) and detected $3,729 \pm 1,189$ transient events per session summed across all segmented neurons. Histological verification confirmed successful viral targeting ($n=8$ animals; Fig. A.3c, Representative verification; Fig. 4.5a) with GRIN lens placements centered over the LHb. For a representative session, the same field of view is shown as a temporal mean projection after MPS preprocessing and window cropping (Fig. 4.5b) and the resulting MPS-extracted spatial footprint map (Fig. 4.5c). Event-aligned, baseline-subtracted activity traces for the most active segmented neurons in that session are shown in Fig. 4.5d. Finally, S2C registered these footprints across sessions, identifying the same LHb neurons recorded on different days: for a representative animal, neurons tracked between two sessions (0.8 mA & Relapse) are shown in green, linked across the two fields of view (Fig. 4.5e).

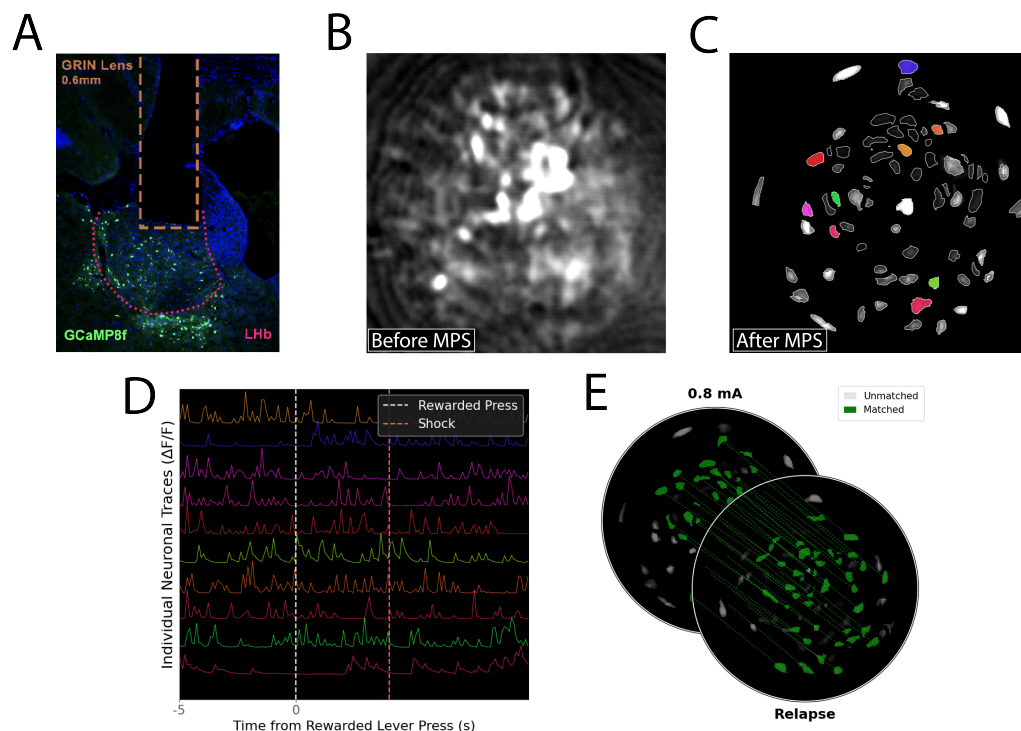


Figure 4.5: Chronic miniscope imaging and cross-session registration of LHB neurons. (A) Representative histology confirming GRIN-lens placement and GCaMP8f expression in the lateral habenula (LHB). Coronal section showing GCaMP8f (green) with DAPI nuclear counterstain (blue); the LHB boundary is outlined (red dashed) and the 0.6 mm GRIN-lens tract overlaid (orange dashed). (B) Temporal mean projection of a representative LHB field of view after MPS preprocessing (background removal, spatial denoising, and motion correction with frame removal) and window cropping. (C) Mean-activity spatial footprint map extracted from the same field of view by MPS, with every segmented neuron outlined in white and the ten most active neurons filled in color. (D) Baseline-subtracted $\Delta F/F$ traces for those ten neurons, aligned to the rewarded lever press (white dashed line) over a window spanning the pre-event baseline through the post-event period, for one representative animal in one session; each trace is colored to match its footprint in (C). (E) Footprint maps from two sessions (top, 0.8 mA; bottom, Relapse) of one representative animal, shown as two circle-cropped fields of view offset in depth. Neurons cross-registered between the two sessions are filled in green and linked by dashed green connectors; unmatched neurons are gray. All panels show one representative animal.

4.3.6 Lhb ensembles shift responsiveness and reorganize as shock intensity rises

For each animal, we found the first take resumed after its within-session strategy shift and aligned Lhb activity to the onset cue, the rewarded lever press, and the shock on that resumed take (Fig. 4.6a–c). The responsive ensembles shifted responsiveness as the shock intensity rose from the onset cue to the shock itself (Fig. 4.6d). At 0.2 mA, the responsive ensembles responded to the onset cue (56% of neurons, separation 0.59) and to the rewarded lever press (45% of neurons, separation 0.60), with no shock ensemble (190 neurons). At 0.8 mA, the responsive ensembles did not respond to the onset cue, but the rewarded lever press (47% of neurons, separation 0.61) and a shock (51% of neurons, separation 0.76; 253 neurons). Lastly, the Relapse session kept the same high-shock layout with no onset cue responsive ensemble, a rewarded lever press responsive ensemble (41% of neurons, separation 0.51), and a shock ensemble (48% of neurons, separation 0.68; 261 neurons). Every retained group survived its own circular-shift null at p below 0.01. In the No Shock session, we found no responsive ensembles to the onset cue or rewarded lever press (separation 0.42 and 0.45, both n.s.; Fig. A.4a). Additionally, the first five rewarded presses of every session, which fall early during loading before any strategy shift, contained no onset cue or rewarded lever press responsive ensembles either (Fig. A.4b–f). Responsive ensembles were therefore present only on takes made under shock after the strategy shift, and absent both when No Shock was scheduled and during the early loading presses that precede any shift.

The signed z-scored response magnitude of the responsive groups shifted from one epoch to the next within each session (Fig. 4.6e). At 0.2 mA, the magnitude fell from the onset cue to the rewarded lever press, from +2.22 to +1.44 (paired Wilcoxon $p = 0.0002$, 138 neurons). At 0.8 mA, it rose from the rewarded lever press to the shock, from +1.50 to +2.01 (paired Wilcoxon $p = 0.008$, 185 neurons). At Relapse, it also rose from the rewarded lever press to the shock, from +1.34 to +1.72 (paired Wilcoxon $p = 0.013$, 183 neurons). Overlap between successive epochs was close to each session’s own responsiveness rate, the fraction expected if the two labels were independent. At 0.2 mA, 54 of 106 (50.9%, 95% CI 41.6–60.3%) cue-responsive neurons were also reward-responsive, against a chance rate of 45.3% (permutation $p = 0.131$). At 0.8 mA, 65 of 120 (54.2%, 95% CI 45.3–62.8%) reward-responsive neurons were also shock-responsive, against a chance rate of 51.4% (permutation $p = 0.237$). At Relapse, 49 of 107 (45.8%, 95% CI 36.7–55.2%) reward-responsive neurons were also shock-responsive, against a chance rate of 47.9% (permutation $p = 0.442$). The same held across days among the neurons tracked and labeled in both sessions. From 0.2 mA to 0.8 mA, 34 of 59 (57.6%, 95% CI 44.9–69.4%) reward-responsive neurons stayed reward-responsive, against a chance rate of 54.5% (permutation $p = 0.829$). From 0.8 mA to Relapse, 26 of 67 (38.8%, 95% CI 28.0–50.8%) reward-responsive neurons stayed reward-responsive, against a chance rate of 36.9% (permutation $p > 0.99$). From 0.8 mA to Relapse, 36 of 66 (54.5%, 95% CI 42.6–66.0%) shock-responsive neurons stayed shock-responsive, against a chance rate of 52.3% (permutation $p = 0.465$; Fig. 4.6f). These confidence intervals are wide. An overlap that does not exceed chance is therefore not evidence that the two labels are independent.

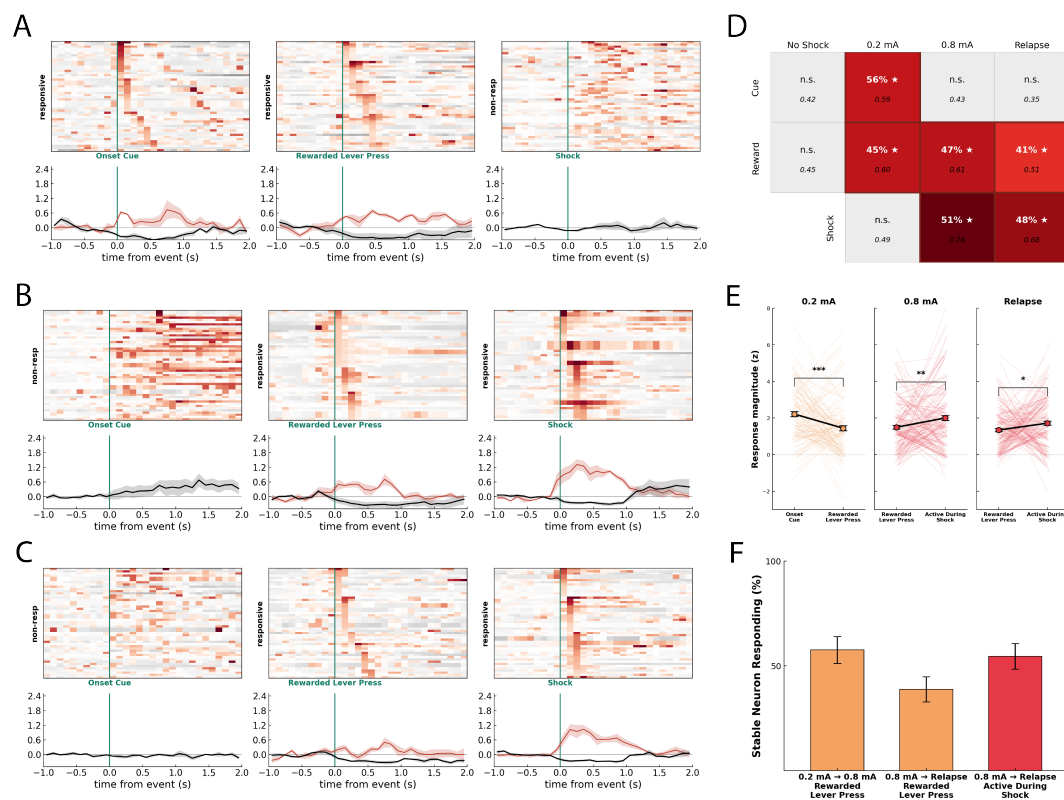


Figure 4.6: LHb ensembles shift responsiveness and reorganize as shock intensity rises. (A) The first rewarded trial each animal resumed after its strategy shift at 0.2 mA, aligned to the onset cue, the rewarded lever press, and the shock. Each panel shows the per-neuron heatmap (top) and the population mean (bottom); neurons are split into responsive and non-responsive groups, and the green line marks the event. (B) The same for 0.8 mA. (C) The same for Relapse. (D) Summary of ensemble recruitment on the resumed take. Rows, event epoch (cue, reward, shock); columns, session (No Shock, 0.2 mA, 0.8 mA, Relapse). Each cell gives the percent of neurons classified responsive to that event, with the peak population separation printed beneath it in italic. Filled cells with a dark border and a star contained a group that survived that panel's circular-shift null (200 permutations, $p < 0.01$); fill intensity scales with the separation. Gray cells marked n.s. did not survive. (E) Signed response magnitude of the responsive group across the two epochs in each session, onset cue versus rewarded lever press at 0.2 mA, and rewarded lever press versus shock at 0.8 mA and at Relapse. Each neuron contributes its z-scored peak response in each epoch window and is paired across the two epochs. Compared by two-sided paired Wilcoxon signed-rank test. $n=138$, 185, and 183 neurons respectively. (F) Cross-day stability of S2C-registered neurons: percent of neurons responsive in the earlier session that remained responsive in the later session. Rewarded-lever-press group from 0.2 to 0.8 mA and from 0.8 mA to Relapse; shock-responsive group from 0.8 mA to Relapse. Tested against the later session's marginal responsiveness rate by within-animal permutation. Data are mean $\pm$ SEM across animals; $n=8$ animals.

4.3.7 The LHB is not predictive, but retrodictive to a Danger press

We found that the shock event was decodable from the LHB population against a time-matched pre-event baseline at 0.2 mA (at-event ROC-AUC 0.67; Fig. 4.7c) and at 0.8 mA (0.88; Fig. 4.7d), but not at Relapse, where it fell to chance at the shock itself (0.33; Fig. 4.7e) despite rising to 0.88 within the following second. Restricting the same decoder to the shock-responsive ensemble rather than to all segmented neurons raised the at-event value at 0.8 mA to 0.93 and at 0.2 mA to 0.78, though at Relapse it remained at chance (0.46), where the shock decode instead peaked 0.3 s after the event. In contrast, neither the onset cue (0.50; Fig. 4.7a) nor the rewarded lever press (0.55; Fig. 4.7b) was decodable above its own time-matched baseline. The press curve does reach 0.62 elsewhere in the window, but at a lag of +1.2 s rather than at the press, and as the best of 41 lags that value is upward-biased by selection rather than evidence of a press-locked response. The population therefore carried the outcome, but not the act of pressing itself. Moreover, by collapsing the same neurons and the same trials into a single bulk trace (similar to fiber photometry) we recovered the same, but weaker dissociation. The shock produced a mean deflection of $+0.18 \pm 0.08$ baseline SD at 0.8 mA (Wilcoxon $p = 0.055$, $n = 8$) and $+0.12 \pm 0.11$ at Relapse ($p = 0.58$, $n = 7$), while the rewarded lever press rose by $+0.02 \pm 0.01$ or less at every shock intensity and the onset cue by less than 0.02 in either direction (all $p > 0.6$).

We next asked whether within-trial activity predicts the next, last, or current rewarded lever press during the Danger period via the same classifier (Fig. 4.7f). A withhold trial’s activity did not forecast whether the animal would press on the next trial (before-miss→press, ROC-AUC 0.50 at 0.2 mA, 0.51 at 0.8 mA, 0.40 at Relapse; Wilcoxon $p \geq 0.47$ in all three), and activity within a press trial, including the shock itself, did not predict whether the animal would withhold on the next trial (in-hit→stop, shock-locked ROC-AUC 0.53; Wilcoxon $p = 0.56$). However, the population retained a decodable trace of the press *after* it had occurred. A withhold trial’s activity did report whether a press had just been made on the preceding trial (after-miss←press), at 0.2 mA (ROC-AUC 0.58; Stouffer $p = 0.043$; $n = 8$), at Relapse (0.64; Wilcoxon $p = 0.047$; $n = 6$), and most strongly at 0.8 mA, where it reached 0.63 across the whole session (Wilcoxon $p = 0.008$; Stouffer $p = 0.016$; $n = 7$) and 0.73 within the Danger phase alone (Wilcoxon $p = 0.031$; Stouffer $p = 0.0015$; $n = 5$). Anchoring the same decoder to the onset cue rather than to trial onset gave the same result (0.66; Stouffer $p = 0.023$).

4.4 Discussion

This study asks whether the LHB signals an adverse consequence during a shift in the strategy of fentanyl self-administration while the drug is still on board, whether that LHB signal runs ahead of the shift or behind it, whether it survives across weeks, and whether the same neurons carry it. We found that: 1) Escalating shock moved the timing of a within-session switch to a low-rate policy and once the shock intensity crossed a threshold the switch became near immediate and persisted a week later. 2) Ultrasonic vocalizations reorganized non-monotonically across the same ladder and

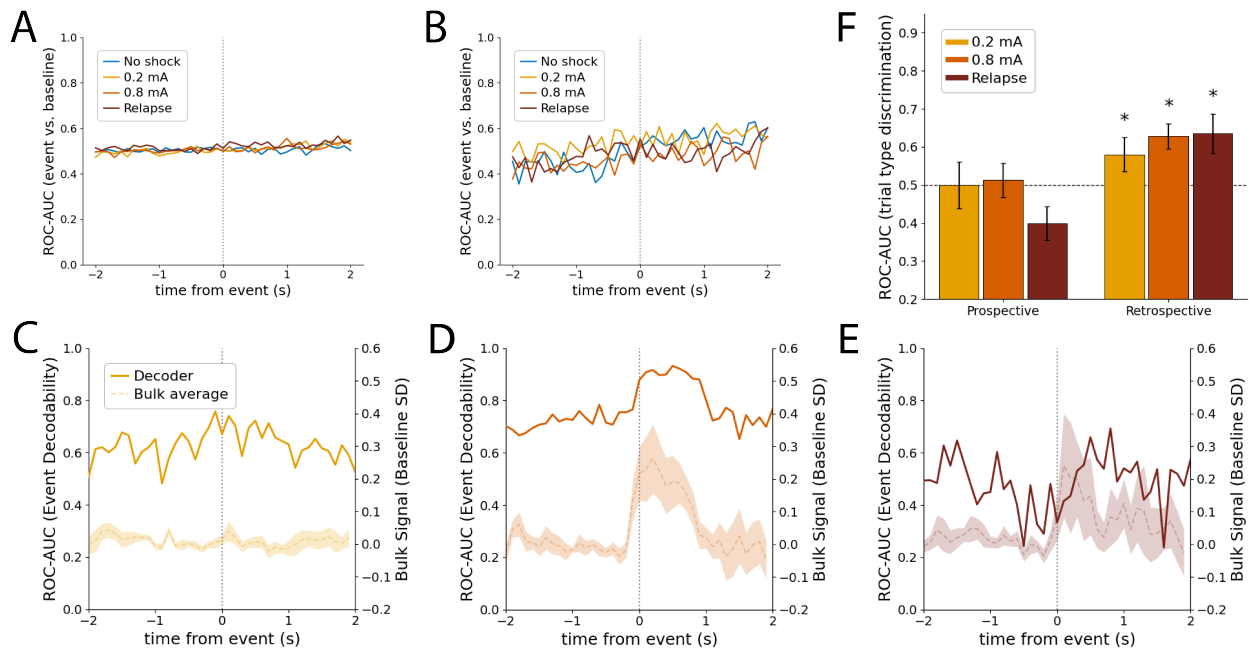


Figure 4.7: The LHB is not predictive, but retrodictive to a Danger press. (A) Time-resolved decodability of the onset cue from the whole LHB population, as out-of-fold ROC-AUC across lags from -2 to $+2$ s relative to the event, decoded against a time-matched pre-event baseline. One line per session (No Shock, 0.2 mA, 0.8 mA, Relapse); dotted line, the event. (B) The same for the rewarded lever press. (C) The same decoder applied to the shock at 0.2 mA, one session per panel. Solid line, left axis: ROC-AUC. Dashed line, right axis: the bulk signal from the same neurons on the same trials, averaged across every neuron in the field of view as a fiber photometry recording would, referenced to the second preceding the event and expressed in baseline SD (mean, shaded band $\pm$ SEM across animals). (D) The same for 0.8 mA. (E) The same for Relapse. (F) Adjacent-trial (sequential) decoding on the isolated-press (miss-hit-miss) motif, all Danger sessions in one panel (bars colored by session: 0.2 mA, 0.8 mA, Relapse). Both decoders read the same withhold trials, at the same anchor, over the same window, and differ only in which neighboring trial is decoded: prospective (before-miss $\rightarrow$ press, does the withhold forecast the upcoming press) and retrospective (after-miss $\leftarrow$ press, does the withhold carry the preceding press). Bars are the mean per-animal ROC-AUC; dashed line, chance. Stars give the Stouffer combination of the per-animal label-shuffle nulls ($*p < 0.05$, $**p < 0.01$). Data are mean $\pm$ SEM across animals; $n=8$ animals per session.

broke at the same intensity, giving a second biological measure of the switch that does not depend on lever pressing. 3) The registered LHB neurons carried the shock while fentanyl was on board, still carried it at Relapse one week later, and reported the press only after the press had happened.

4.4.1 Conventional shock intensities bracket the behavioral transition

We ran the same rats through the intensities that the field generally dichotomizes. At 0.4 mA, the canonical self-administration intensity, Danger-phase responding was statistically indistinguishable from the Days 1-19 baseline (18.8 ± 2.6 rewarded lever presses, $W = 112$, $p = 0.17$), and nothing across 0.1 to 0.4 mA graded with intensity (Friedman $\chi^2(3) = 1.94$, $p = 0.59$). At 0.8 mA, the canonical stress intensity, the same animals nearly stopped (1.3 ± 0.5 presses, $W = 0$, $p < 0.001$) and remained suppressed a week later, with the press rate they retained after their shift falling from $5.9 \pm 1.0\%$ to $3.3 \pm 0.5\%$ of their own pre-shift rate. Thus, in our paradigm, a study that ended at the mild convention would have called this fentanyl self-administration shock resistant, whereas a study that used only the stress convention would have called it shock sensitive. In either case, both readings would have been artifacts of an inherited number rather than facts about the animal. Neither convention was ever anchored to habenular physiology, since acute recordings of LHB shock encoding have used 0.3 mA (Lecca et al., 2017), and no study has recorded the same LHB neurons across a range of intensities.

Reading that collapse as a decision assumes the animals could still press and that a fixed current meant the same thing to them that it means to a drug-free animal. Safe-trial responding did not change across any of the six sessions ($\chi^2(5) = 8.32$, $p = 0.14$) while Danger-trial responding collapsed, so within the same session at the same drug load in the same animal the brain answered each contingency differently. Movement rose rather than fell during the Danger phase (2243 ± 29 cm/trial at 0.8 mA versus 1985 ± 24 cm/trial at No Shock, $F(3,5481) = 18.03$, $p < 0.001$), so the refusal was a decision and not an incapacity. The second assumption does not survive, and the design does not require it to. Fentanyl was pharmacologically active across the entire Danger window, at 58 to 88% of each animal's own baseline peak at Danger onset and at or above half that peak for the following 8 min. Analgesia raises the current an animal will tolerate, so the 0.8 mA that suppressed responding here was, in the terms a drug-free animal would experience it, milder than the 0.8 mA the stress literature calibrated. That is the design rather than a confound, since the state we set out to record is the one in which the drug is already working, and what the design establishes is that the discrimination survives the drug.

Mirroring treatment literature organized around abstinence, most preclinical models are built almost entirely around choices made with no drug on board and cognitive control intact (Marchant et al., 2013; Fredriksson et al., 2021c; Pelloux et al., 2007). However, a framework organized this way never asks whether a mechanism for stopping mid-episode can be learned at all, because it never puts the animal in the state where such a mechanism would be used. The animals here learned such a mechanism, in that they kept responding on Safe trials and withheld on Danger trials while fentanyl was pharmacologically active in both cases. An adverse consequence is what

installed that capacity in this paradigm rather than the only thing that could install it, and what the result establishes is that a mechanism for stopping mid-episode exists and can be trained in the state in which it would be used. Therefore, before moving into the clinic, whether an appetitive alternative reaches the same policy is the next step, given that food and social choice procedures produce voluntary abstinence in drug-free animals (Venniro et al., 2020; Fredriksson et al., 2021c).

4.4.2 Ultrasonic vocalizations dissociate affective state from behavioral output

Vocalizations gave a second biological readout of the same ladder. Under No Shock the pattern was the expected one with high-frequency calls concentrated in the loading phase (59.1 ± 25.7 calls per animal over 0-15 min) and almost no low-frequency calls thereafter (Coffey et al., 2023). At 0.1 mA, the intensity where responding trended upward, a second high-frequency peak appeared late in the session, which would be consistent with a mild shock acquiring some appetitive weight. Conversely, the late peak vanished at 0.2 mA and calling collapsed almost entirely at 0.4 mA (3.8 ± 2.9 versus 80.8 ± 32.4 calls per animal, $p = 0.0078$), while Danger responding was unchanged at both. Affect and behavior therefore came apart, and the simplest interpretation of the 0.1 mA peak is a response to novelty rather than to value. However, during the intensity where responding dramatically decreased, the peak then returned at 0.8 mA arriving 319 ± 18 s after Danger onset and followed by a large low-frequency surge (13.1 versus 0.1 calls per animal over 30-60 min) dominated by Long 22-kHz calls (29.0 ± 29.0 calls per animal over 25-40 min).

Deductively, a signal that appears when the animal starts responding more and again when it stops responding altogether cannot be tracking valence in any simple sense. Thus, we argue that this is better understood as marking a change in what the animal is doing. For example, at Relapse the second high-frequency peak was gone and the low-frequency pattern was bimodal (17.2 calls early, 37.4 calls late, a 2.17-fold ratio). Animals switched to the low-rate policy within 2.2 ± 0.6 trials of Danger onset at Relapse, against 12.2 ± 4.9 trials at 0.8 mA. In line with the simple negative affect interpretation (Koob, 2021), low-frequency calling rose at the two intensities where responding collapsed, from 0.1 calls per animal over 30-60 min at No Shock to 13.1 at 0.8 mA and 37.4 at Relapse, with total calling rising from 3.8 ± 2.9 to 140.6 ± 100.7 calls per animal between 0.4 and 0.8 mA ($p = 0.0312$). The omnibus test across sessions did not reach significance (Friedman $\chi^2(5) = 10.25$, $p = 0.069$), reflecting substantial individual variability in low-frequency call production. However, calls cannot be locked to individual trials in this paradigm since vocalizations are emitted in bouts that reflect a sustained internal state rather than one call per event, and a bout initiated on one trial continues across the trials that follow it at latencies that are themselves variable and long relative to a trial. USVs are therefore a session-level and epoch-level measure of affective state, which is the level at which we use them here.

4.4.3 Shock modulates the timing of a within-session policy shift

Animals at every intensity kept responding for a variable number of Danger trials and then dropped abruptly, so a press rate or a cumulative curve would have averaged that structure away. The

changepoint model resolved a shift at the 5% level in the large majority of animal-sessions in every condition, from 81% under No Shock to 96% at Relapse. Since a shift was resolved in most sessions even under No Shock, and animals retained roughly a fifth of their pre-shift rate when no shock was scheduled ($18.5 \pm 2.0\%$), the drop after the shift is best understood as the pharmacological transition from loading to maintenance rather than as a consequence of the shock, which cut into that retained rate only once its intensity became suppressive. We found that shock affected the timing of this shift non-monotonically, rising from 9.2 ± 3.1 trials after Danger onset at No Shock to 49.2 ± 9.9 trials at 0.2 mA, falling to 30.2 ± 8.2 at 0.4 mA and 12.2 ± 4.9 at 0.8 mA, then collapsing to 2.2 ± 0.6 at Relapse. Both boundary slopes were individually signed ($+34.32$ and -33.68 , both $p < 0.0001$) with an interior turning point between 0.2 and 0.4 mA (joint $p < 0.0001$). Paralleling the escalation at 0.1 mA and prior work on stressor intensity and drug reinforcement, moderate intensities delayed the switch rather than advancing it (Shaham and Stewart, 1994; Miczek and Mutschler, 1996; Piazza and Le Moal, 1998; Han et al., 2015).

Nonetheless, we found that both biological measures therefore describe the same ladder in the same shape. For instance, vocal output and shift timing were each non-monotonic across intensities, each was preserved through the mild rungs, and each broke at 0.8 mA. Moreover, given that high-frequency calls preceded Danger onset by 149 ± 170 s while the strategy changepoint followed it by 1009 ± 303 s in the same animal-sessions ($W = 67$, $p = 0.0007$), then the calling tracked the appetitive loading phase rather than the switch and cannot be read as a moment-by-moment marker of when the policy changed. Therefore, we conclude that USVs are an independent, non-behavioral measure that reorganizes at the intensity where the policy reorganizes.

4.4.4 Multi-hour recording and cross-session registration require dedicated computational tools

This design imposes two requirements that existing tools do not meet. The shock has to land while the drug is on board, which forces multi-hour sessions, and each session ran 170 ± 5 min for $101,721 \pm 2,705$ frames at 10 Hz, across 30 sessions and 2,027 neuron-sessions. Standard pipelines do not scale to that length, which is the limitation MPS was built to remove (Peden-Asarch et al., 2025), and segmentation yield here was stable across the imaging window (Spearman $\rho = -0.28$, $p = 0.14$, session $F(4,18) = 0.65$, $p = 0.64$) rather than degrading with session duration. The second requirement was with regard to the fact that asking whether the signal survives a week is not the same as asking whether the LHB responds a week later. Furthermore, since the field of view shifts between mountings and the scope may be remounted in a different orientation, with 5 of 15 adjacent-session registrations here related by rotations exceeding 90° , holding neuronal identity across days required S2C (Peden-Asarch et al., 2026e), which recovered those transforms by RANSAC without manual reorientation. Both tools are open source, and since any paradigm that pairs multi-hour sessions with imaging across weeks faces the same two requirements, we released them for that use rather than as an accessory to this dataset.

4.4.5 LHB ensemble reorganization coincides with the behavioral threshold

We aligned LHB activity to the first take each animal resumed after its own strategy shift, so the neural question was asked on the trial the behavior had already identified. Ensembles were present only on takes made under shock after the shift. Under No Shock we recovered no ensemble to the onset cue or the rewarded lever press (separation 0.42 and 0.45, both n.s.), and the first five presses of every session, which fall during loading before any shift, contained none either. That absence agrees with fiber photometry reporting no LHB response to unshocked oral fentanyl self-administration (Coffey et al., 2023) and it means the ensembles shown are specific to fentanyl self-administration under adverse consequences rather than to fentanyl self-administration alone.

At 0.2 mA, where responding was at baseline, the population organized around the onset cue (56% of neurons, separation 0.59) and the rewarded lever press (45%, 0.60) and formed no shock ensemble at all. At 0.8 mA, where responding collapsed, the cue ensemble was gone and the population organized around the press (47%, 0.61) and the shock (51%, 0.76). Relapse kept the high-intensity layout a week later (press 41%, 0.51, shock 48%, 0.68). Notably, the presence of a press-locked ensemble in LHB during oral fentanyl self-administration has not, to our knowledge, been reported, and would have been inaccessible to a bulk signal. Additionally, since the press ensemble and the shock ensemble were resolved in the same neurons on the same trial, we recovered a drug-axis representation and a shock-axis representation in one population in one paradigm.

Additionally, the shock was decodable from the whole population at 0.2 mA at an ROC-AUC of 0.67, well above chance, rising to 0.88 at 0.8 mA and 0.93 when restricted to the shock ensemble. A shock that did not change behavior was therefore not neurally absent, in that it remained decodable from the population at 0.2 mA while recruiting no coherent ensemble. Thus, the behavioral shift coincides with the organization of a synchronized group of neurons and not the arrival of a shock signal. Therefore, we suggest that this population is not the representation of a shock signal but the recruitment of a synchronized ensemble carrying it. Interestingly, the same recordings collapsed to a bulk trace reproduce the published photometry picture almost exactly, with the shock moving the average by $+0.18 \pm 0.08$ baseline SD at 0.8 mA ($p = 0.055$) and the press and cue moving it not at all. The difference between 0.18 and 0.88 is not resolution but weighting. A fiber sums every neuron equally and this signal does not survive that sum. Thus, whether an aggregate suffices is a fact about a given structure and a given signal, and it is only decodable from data that keep what the average discards (Legaria et al., 2022; Yun and Parker, 2025). For a nucleus with segregated afferents (Hu et al., 2020; Barker et al., 2017), split efferents (Hikosaka et al., 2008), and multiple molecular subtypes (Hashikawa et al., 2020; Wallace et al., 2020), a modest photometry number (Coffey et al., 2023) cannot be its only interpretation.

4.4.6 LHB activity is retrodictive rather than predictive of Danger presses

If the LHB gates the choice to press a lever despite adverse consequences, then its activity would represent that choice before the choice is made. A withhold trial’s activity did not forecast whether the animal would press on the next trial (0.50 at 0.2 mA, 0.51 at 0.8 mA, 0.40 at Relapse, all

$p \geq 0.47$), and activity within a press trial, including the shock itself, did not predict whether the animal would withhold on the next one (0.53, $p = 0.56$). Thus, given that the shock is the most decodable event in the session and it forecasts nothing that follows it, then it is unlikely that the LHB is gating the choice to press a lever despite adverse consequences during fentanyl self-administration. However, we found that the LHB carries the press after the fact (retrodictively), reaching a per-animal ROC-AUC of 0.73 within the Danger phase at 0.8 mA (Wilcoxon $p = 0.031$, Stouffer $p = 0.0015$), 0.63 across that whole session, and holding at 0.2 mA and at Relapse.

One interpretation is that a failure to decode is ordinarily uninformative, since it can be attributed to too few neurons, too few trials, or too blunt a classifier. However, here the prospective and retrospective decoders ran on the same neurons, the same trials, and the same cross-validation, and differ only in whether they look forward or back. Therefore, although the LHB is usually thought of as an aversion signal that drives suppression, we found that it sits downstream of the choice to press a lever despite adverse consequences during fentanyl self-administration.

Given that roughly half the press-responsive neurons at 0.8 mA were also shock-responsive, and roughly half the shock-responsive neurons stayed shock-responsive into Relapse, then this points to this retrodictive signal sitting in a single brain area. For instance, if a stable half of the population carries an outcome signal that the LHB did not compute, then the most economical source is a single area, with the remainder recruited from elsewhere. Prefrontal projections to LHB are positioned to carry outcome information and behavioral control (Warden et al., 2012), which is the content this signal has, and mPFC is where we suggest to look first. As further evidence, entopeduncular inputs signal omitted reward and negative prediction error (Stephenson-Jones et al., 2016), lateral preoptic inputs are recruited by noxious stimuli and their predictive cues (Barker et al., 2017), and lateral hypothalamic inputs convey homeostatic and motivational state (Stamatakis et al., 2016), any of which could account for the rest. Retrograde labeling or projection-specific expression combined with the imaging is what would test it.

4.4.7 Limitations

Cross-session registration recovers only the subset of neurons visible on multiple days, so claims about the stability of an ensemble apply to that subset and not to the LHB as a whole. Imaging was restricted to Days 19, 21, 23, and 30, so the LHB was recorded at 0.2 mA and at 0.8 mA but not at 0.4 mA, the highest intensity at which Danger responding was still preserved. These recordings therefore bracket the behavioral transition rather than localize it, and the question of whether LHB encoding changes at the same rung the behavior does, rather than somewhere between 0.2 and 0.8 mA, is one the imaging cannot settle. The trajectory between the imaged sessions is likewise inferred rather than observed, and we recorded without manipulating, so the relationship between any LHB signal and the decision to stop remains correlational. In that case, if the LHB reports the press after the fact, then silencing or driving it tests whether downstream readers use that report rather than whether the LHB computes the decision, and a null result would be uninformative about the question we asked. What the retrodictive result points to instead is a manipulation upstream in the

afferents that supply the signal, or a closed-loop intervention timed to each animal’s own strategy shift rather than to the session. The ladder was also escalated within animals in a fixed ascending order, so the effect of an intensity is not separable from the history of intensities preceding it, and prior exposure to high shock is known to make self-administration more responsive to an intensity that was previously ineffective (Durand et al., 2022). Finally, we did not measure flinch or jump thresholds, so we cannot rule out that some of the between-animal variability in Fig. 4.2e reflects nociceptive sensitivity rather than a decision process, although shock insensitivity in rats has been traced to a failure of instrumental control rather than to blunted aversion (Jean-Richard-dit Bressel et al., 2019).

4.4.8 Implications for decision-making under active drug exposure

The clinically relevant scenario this paradigm models is the one in which an individual has already consumed, is experiencing the drug’s effects along with the associated impairment and elevated motivation, and must decide whether to continue despite knowing what may follow. We found that the decision to stop depended on a specific consequence being present rather than on stable anti-drug learning, and thus, imply that individuals who achieve cessation through legal sanctions, medical crises, or relationship ultimatums may therefore remain vulnerable in contexts where those consequences seem absent or distant.

These animals acquired a policy for stopping while the drug was on board, which is the capacity an abstinence-first framework has no way to train because it never puts the individual in that state, and whether the same policy can be reached through a competing reinforcer or an active coping alternative is the test that has to come before any of this reaches a clinic (Venniro et al., 2020; Fredriksson et al., 2021c). Additionally, we found that the adverse consequence was not forgotten across the abstinence week, since the shock ensemble at Relapse was recovered at a strength comparable to 0.8 mA, and thus we suggest that treatment does not need to reinstall the knowledge that using carries a cost so much as keep that cost engaged at the moment the choice is made.

Data availability

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The Miniscope Processing Suite (MPS) is available at https://github.com/ariasarch/MPS_1.0.0. Stars2Cells is available at https://github.com/ariasarch/Stars2Cells_GUI. Analysis code is available from the corresponding author upon reasonable request.

Acknowledgments

Funding for this work and for the authors' research effort was provided by the National Institute on Drug Abuse (T32DA007278, R01DA052618, and R00DA052571) and a pilot grant from the Puget Sound VA Research and Development Service Line.

Appendix A

Supplementary material for Chapter 4

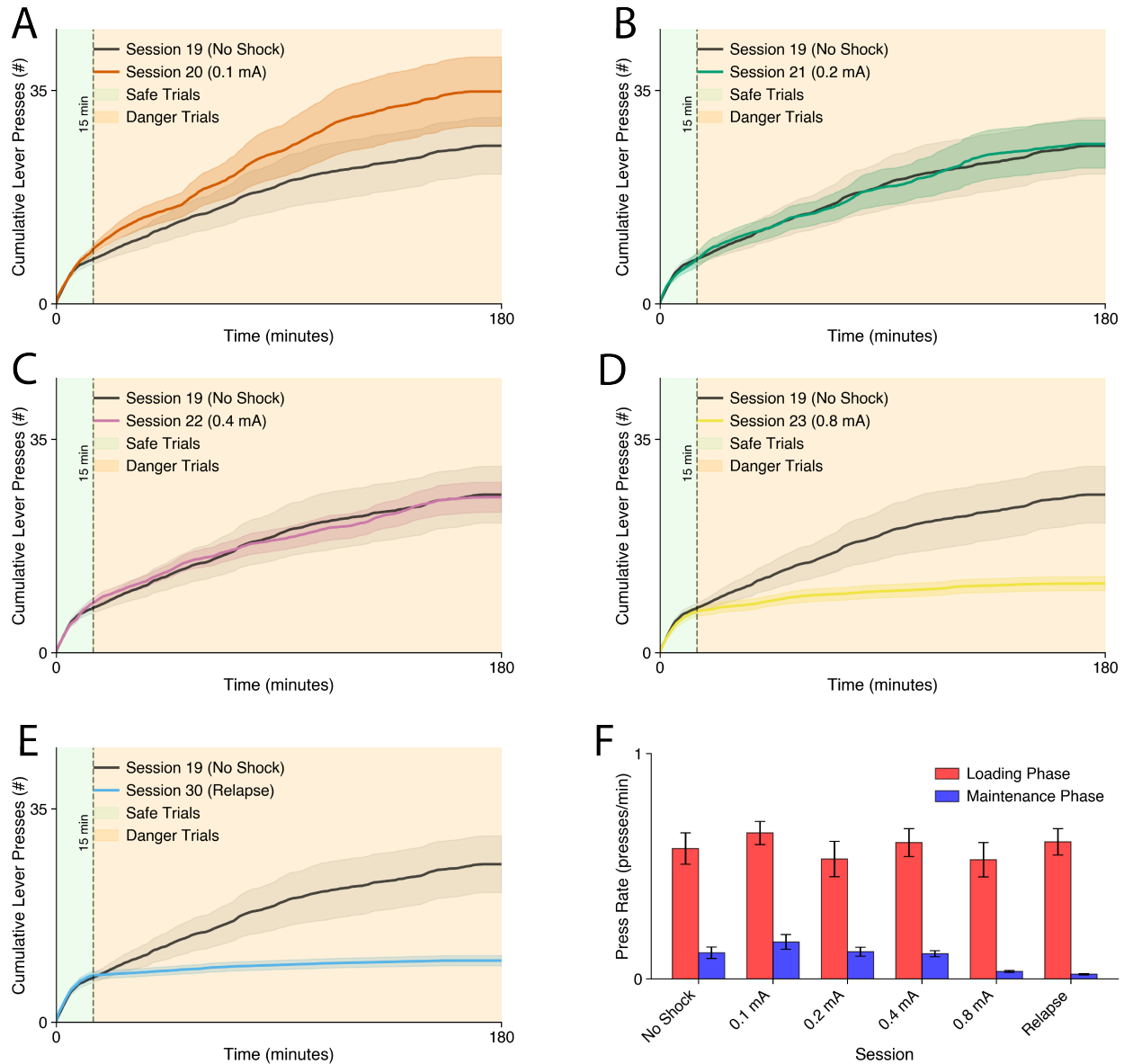


Figure A.1: Per-session distributions and phase-resolved press rates underlying shock-induced suppression. (A) Danger-trial responding at 0.1 mA (Day 20) relative to the Day 19 baseline. (B) The same for 0.2 mA (Day 21). (C) The same for 0.4 mA (Day 22). (D) The same for 0.8 mA (Day 23). (E) The same for the Relapse (Day 30, 0.8 mA). (F) Rewarded lever press counts during the loading phase (0 to 12.5 min) and the maintenance phase (12.5 to 180 min) for each shock session, each compared with the Days 1–19 baseline by Mann–Whitney U test. Data are mean $\pm$ SEM across animals; $n=27$ animals.

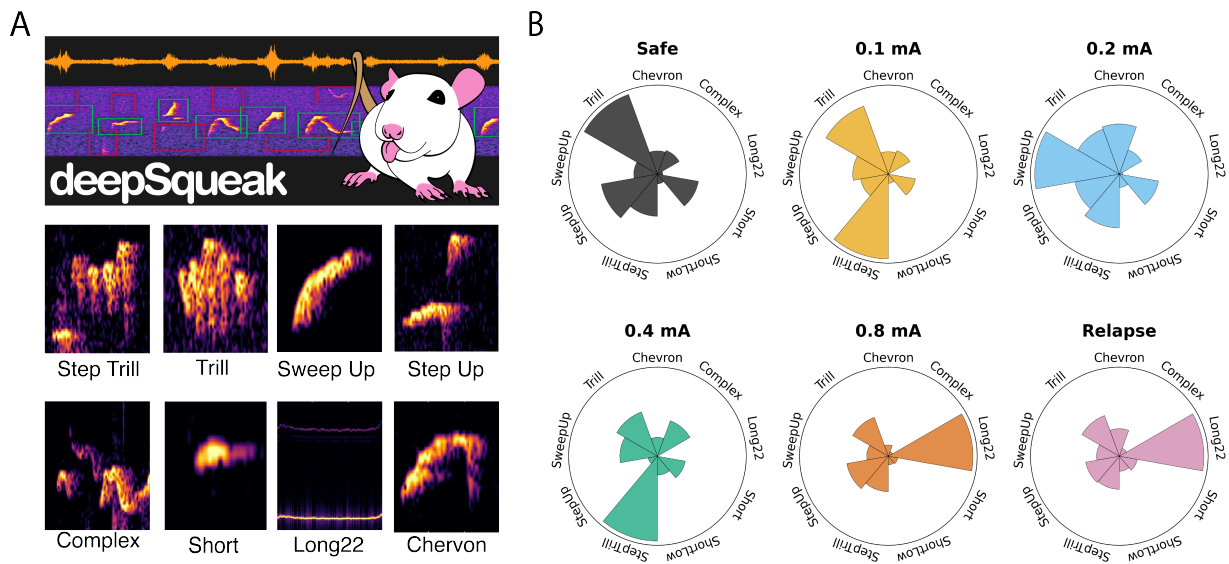


Figure A.2: Unsupervised classification of ultrasonic call types and their redistribution across shock intensities. (A) Representative spectrograms of the nine call types identified by DeepSqueak, one example per type (Chevron, Complex, Long22, Short, ShortLow, StepTrill, StepUp, SweepUp, Trill). (B) Rose plots of call type composition at each shock intensity. Each wedge is one call type and wedge length gives the total number of calls of that type, pooled across animals. Wedge length is a total call count pooled across animals, not a per animal mean. $n=8$ animals.

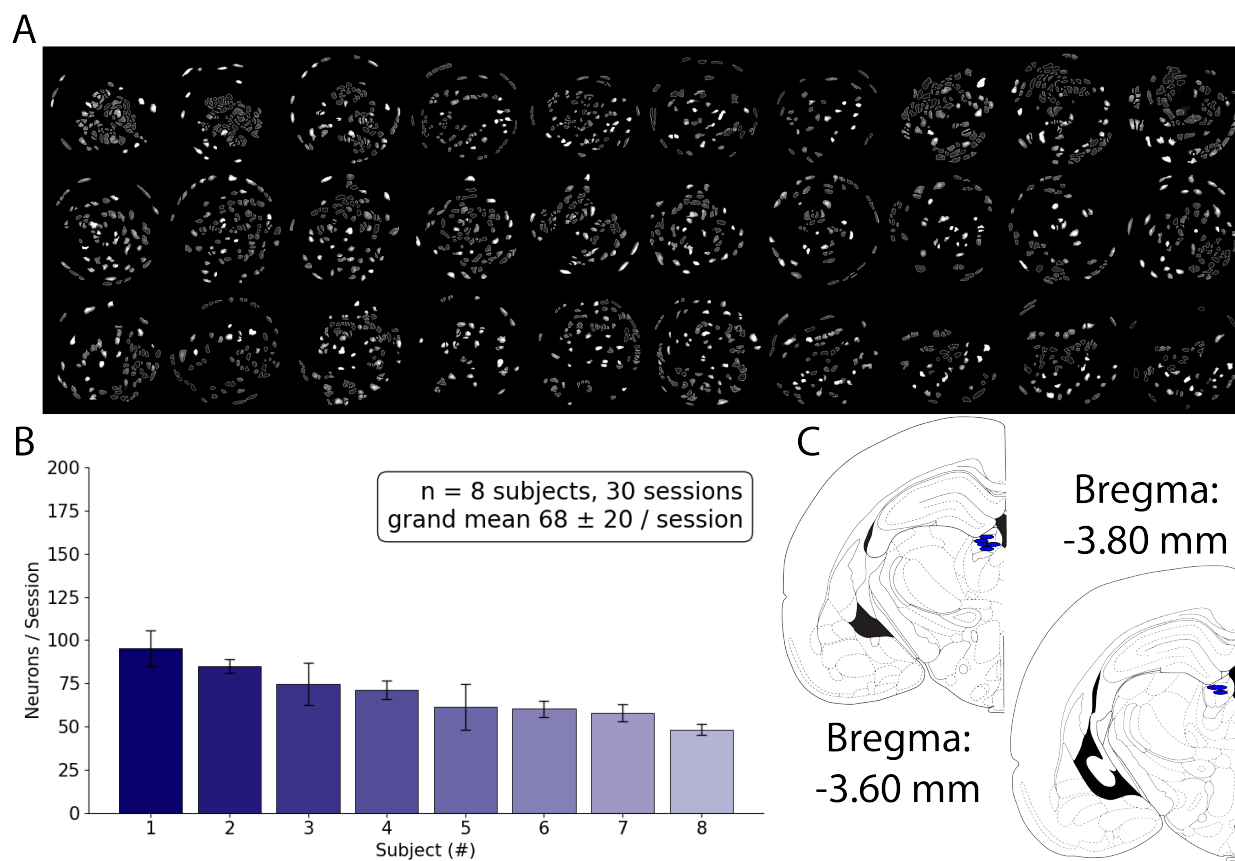


Figure A.3: Segmentation yield, per-session count stability, and GRIN-lens targeting across the imaged cohort. (A) Contact sheet of MPS-extracted spatial footprint maps (A-maps) for all imaged sessions. Each panel shows the activity-weighted footprint sum (grayscale) with per-cell contours drawn at 30% of each footprint's peak; panels are ordered by animal and imaging day. $n=30$ sessions from 8 animals. (B) Segmented neurons per session, by animal. Bars, mean $\pm$ SD across that animal's sessions; dots, individual sessions. (C) Reconstructed GRIN-lens placements for all 8 animals (colored markers, one per animal) overlaid on a coronal schematic at the level of the lateral habenula (LHb).

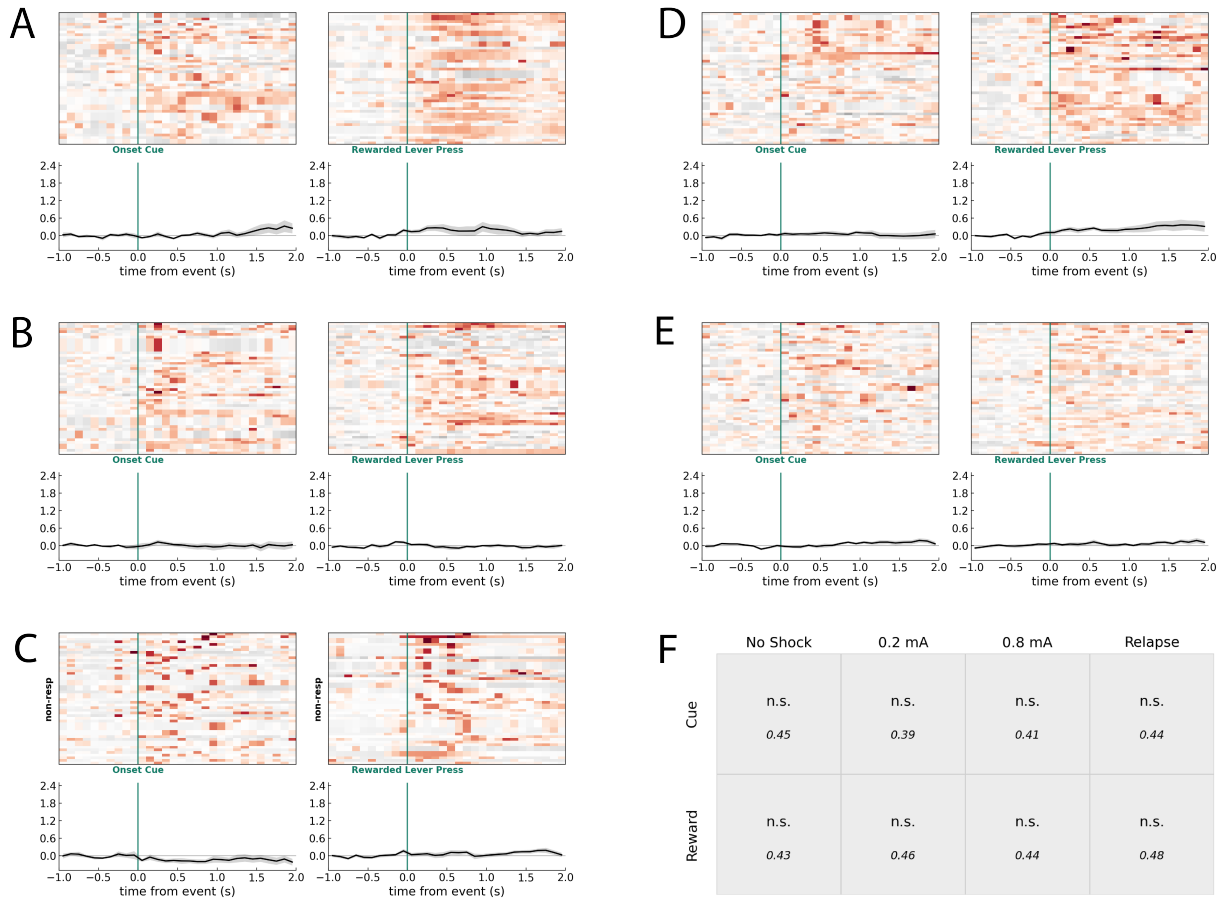


Figure A.4: No LHB ensembles emerge during the rewarded Safe trials or the No Shock session. (A) The first rewarded trial after the strategy shift in the No Shock session, aligned to the onset cue and the rewarded lever press. Each panel shows the per-neuron heatmap (top) and the population mean (bottom); neurons are split into responsive and non-responsive groups, and the green line marks the event. (B) The first five rewarded presses of the No Shock session, averaged within animal and aligned to the onset cue and the rewarded lever press, in the same format as (A). (C) The same for 0.2 mA. (D) The same for 0.8 mA. (E) The same for Relapse. (F) First-press summary across sessions, in the same format as Fig. 4.6d. Bands were classified on the odd presses and tested on the held-out even presses against a circular-shift null. Data are mean $\pm$ SEM across animals; n=8 animals.

Chapter 5

General Discussion

5.1 Summary of findings

This dissertation investigated the timing of lateral habenula dynamics while an animal keeps self-administering fentanyl with the drug on board despite adverse consequences and whether those dynamics are retained across weeks by the same neurons. To answer this, we needed the ability to process recordings long enough to capture the sessions of interest and register neurons confidently across weeks, and neither existed. Thus, here we built and released the Miniscope Processing Suite, which carried 7.26 TB and 77 h of raw video through source extraction in 55.6 h on a single workstation and recovered its components against synthetic ground truth. We then built and released Stars2Cells, which registered neurons from four-neuron geometry at an F1 of 98.4% and held 41.6% of DMS neurons across four sessions against the standard. As further validation of the utility of these two tools, we show that in the dorsomedial striatum, a press response conserved across a fentanyl demand curve while roughly half of the neurons switched class at every transition and was rebuilt from a different set of neurons each session. As for our overarching question, behaviorally, we found that shock moved the timing of a within-session switch from loading to maintenance and that ultrasonic vocalizations reorganized across the same ladder while dissociating from responding. Using this shift in strategy and via the two new tools, we looked at the LHb and found that it reported the shock while the drug was on board and still reported it at Relapse a week later.

5.2 The Miniscope Processing Suite removes the scale limit

The first requirement was to record long enough multi-hour sessions that standard pipelines do not complete, since they were built for 15 to 30 min recordings and fail on datasets of this duration for want of memory (Giovannucci et al., 2019b; Zhou et al., 2018a; Lu et al., 2018; Dong et al., 2022a). Across 30 sessions totaling 7.26 TB and 77 h of raw video, with sessions running up to 2.9 h, MPS carried every stage from preprocessing through source extraction and deconvolution to completion in 55.6 h on a single workstation, at a processing rate of 0.7 to 0.9 min per minute of recording. Since it runs from a point-and-click interface with no code, the pipeline is usable without

the computational background that existing tools assume, which is the second barrier that has kept miniscope analysis narrow. Additionally, because no cell-resolved ground truth exists for a real recording, MPS was validated on a fully synthetic dataset built to match the acquisition parameters of the real data, with 100 neurons in a 600 by 600 pixel field of view at 10 Hz over a 3 h session, then degraded with neuropil haze, a vignette, sensor noise, and motion. Run on the degraded video at 10, 20, 60, and 180 min, MPS recovered all 100 ground-truth neurons at every duration (precision and recall of 1.0, $F1 = 1.000$), reaching a mean per-neuron r^2 of 0.987 between recovered and true traces and a mean spatial area ratio of 0.984, and holding at 0.982 and 0.911 in the full 180 min session. Since every existing pipeline fails at the recording length used here, no fair head-to-head comparison on these data is possible, which is itself the gap MPS was built to close.

5.3 Stars2Cells recovers same-cell identity across sessions

The second requirement was to hold the same neurons across weeks, which the mainstream registration approach cannot confidently do, since it erroneously matches regions of interest by footprint overlap, centroid distance, or signal correlation (Sheintuch et al., 2017b; Johnston et al., 2022; Majnik et al., 2025; Vergara et al., 2025a; Hattori and Komiyama, 2022; Lagache et al., 2021). S2C solves this by mapping each neuron by the geometry of its four-neuron neighborhood, matching those descriptors under rotation, translation, and uniform scaling, verifying those correspondences with RANSAC, and assigning them by Hungarian matching (Hogg et al., 2008; Lang et al., 2010). On a synthetic benchmark of 1,262 paired runs with construction-time ground truth, S2C reached a mean F1 of 98.4% at a precision of 99.9% and a recall of 97.2%, and it exceeded ROI-based matching at every neuron count and on every condition tested. Two independent tests establish that this accuracy transfers off the synthetic data. Run blind on 431 Allen Brain Observatory two-photon containers, S2C reproduced the Allen Institute’s own cross-session registration at a median F1 of 0.97. On our own chronic one-photon recordings, S2C tracked DMS neurons through a four-session registration chain sampling the demand curve at 222, 70, 22, and 0 $\mu\text{g/mL}$, matching $69.1 \pm 7.2\%$ of neurons at the first transition against $13.0 \pm 3.3\%$ for footprint matching and retaining $41.6 \pm 3.4\%$ of first-session neurons end to end against $13.0 \pm 3.9\%$, a 3.2-fold improvement in longitudinal yield.

5.4 The lateral habenula runs behind the choice

A within-session changepoint model made this switch measurable trial by trial, and it recovered a shift in the large majority of animal-sessions in every condition, including under No Shock. Since the switch was present even when no shock was scheduled, it is best read as the pharmacological transition from loading to maintenance that the drug imposes on its own. That timing traced an inverted-U across the shock ladder, rising from 9.2 ± 3.1 trials after Danger onset under No Shock to 49.2 ± 9.9 at 0.2 mA, falling to 30.2 ± 8.2 at 0.4 mA and 12.2 ± 4.9 at 0.8 mA, then collapsing

to 2.2 ± 0.6 at Relapse, with both boundary slopes individually signed and an interior turning point between 0.2 and 0.4 mA. A mild adverse consequence therefore delayed the switch while a suppressive one advanced it, so that the same behavioral read-out moved in opposite directions depending on intensity and a session total would have averaged both away. Once the switch occurred, the rate the animal retained afterward stayed near a fifth of its pre-shift rate through the mild rungs and fell only when intensity became suppressive, at 14.6% across low intensities against 4.6% across high, so shock cut into the maintenance rate only once it was strong enough to suppress.

Ultrasonic vocalizations gave a second biological read-out that reorganized across the same ladder while coming apart from behavior. Positive-affect high-frequency calling collapsed at 0.4 mA, the intensity at which Danger responding was still at baseline, so affect fell while behavior did not, and negative-affect Long 22-kHz calling rose only at the two intensities where responding collapsed. This means that the switch to withholding is accompanied by a shift in affective state that a press count alone would miss. However, what the vocalizations do not provide is a trial-by-trial marker, since high-frequency calls preceded Danger onset by 149 ± 170 s while the strategy changepoint followed it by 1009 ± 303 s in the same animal-sessions, so the calling tracked the appetitive loading phase and indexes affective state at the session and epoch level rather than moment by moment.

Holding the shock 4 s behind the rewarded press also meant that a single population could be read along both axes the LHb has been assigned in two literatures that had not previously met, namely the drug-axis signal that tracks the value of the drug itself (Coffey et al., 2023) and the shock-axis signal that tracks an externally imposed adverse consequence (Matsumoto and Hikosaka, 2009; Wang et al., 2017; Congiu et al., 2019; Shabel et al., 2019). At 0.2 mA, where responding was at baseline and the drug axis was undisturbed, the population organized around the onset cue (56% of neurons, separation 0.59) and the rewarded lever press (45%, separation 0.60) and formed no shock ensemble at all, whereas under No Shock neither the cue nor the reward recruited an ensemble (separation 0.42 and 0.45, both n.s.). As the adverse consequence intensified to 0.8 mA the cue ensemble dropped out and the population reorganized around the press (47%, 0.61) and the shock (51%, 0.76), so that the drug axis and the shock axis were carried not by two separate systems but by one population that handed a conserved press ensemble from a cue-dominated layout to a shock-dominated one as the intensity rose.

As for whether this was specific to the presentation of an adverse consequence, Safe-trial responding did not change across any of the six sessions ($\chi^2(5) = 8.32$, $p = 0.14$) while Danger-trial responding collapsed from 18.8 ± 2.6 rewarded presses at 0.4 mA to 1.3 ± 0.5 at 0.8 mA ($W = 0$, $p < 0.001$), so that within the same session at the same drug load in the same animal the brain answered each contingency differently. Movement also rose rather than fell during the Danger phase (2243 ± 29 cm/trial at 0.8 mA versus 1985 ± 24 cm/trial under No Shock, $F(3,5481) = 18.03$, $p < 0.001$), and the latency from lever to port on the trials where presses occurred did not increase, so we can infer that the refusal was a decision rather than an incapacity to respond. It is important to highlight that fentanyl was pharmacologically active across the entire Danger window, at 58 to 88% of each animal's own baseline peak at Danger onset and at or above half that peak for the following

8 min, and since opioid-induced analgesia raises the current an animal will tolerate, the 0.8 mA that suppressed responding here was, in the terms a drug-free animal would experience it, milder than the 0.8 mA the stress literature calibrated. Since the state we set out to record is the one in which the drug is already working, what this design establishes is that the discrimination survives the drug being on board.

Notably, we found that the shock was the most decodable event in the session, reaching an ROC-AUC of 0.88 from the whole population at 0.8 mA and 0.93 within the shock ensemble, and yet it forecast nothing that followed it (0.53, $p = 0.56$ for whether a press trial predicts a withhold on the next one). A withhold trial did not forecast whether the animal would press on the next trial either (0.50 at 0.2 mA, 0.51 at 0.8 mA, 0.40 at Relapse, all $p \geq 0.47$), whereas the same withhold trial carried the press that preceded it at 0.73 within the Danger phase at 0.8 mA. A failure to decode is ordinarily uninformative, since it can be attributed to too few neurons, too few trials, or too blunt a classifier, but here the prospective and retrospective decoders ran on the same neurons, the same trials, and the same cross-validation, and differed only in whether they looked forward or back. On that same note, the ladder also separates a decodable signal from a behaviorally effective one, which the field routinely conflates. At 0.2 mA the shock was still decodable from the whole population at an ROC-AUC of 0.67, well above chance, even though it recruited no shock ensemble, so that a shock producing no significant behavioral change at 0.2 mA was not thereby neurally absent. In addition, at 0.4 mA, the canonical self-administration intensity, Danger-phase responding was statistically indistinguishable from baseline, and nothing across 0.1 to 0.4 mA graded with intensity (Friedman $\chi^2(3) = 1.94$, $p = 0.59$), whereas at 0.8 mA, the canonical stress intensity, the same animals nearly stopped and stayed suppressed a week later at Relapse. A study that ended at the mild convention would have called this fentanyl self-administration shock resistant, and a study that used only the stress convention would have called it shock sensitive, so that both readings would have been artifacts of an inherited number rather than facts about the animal.

Thus, although the LHb is usually thought of as an aversion signal that drives suppression, we found that it sat downstream of the choice to press the lever despite an adverse consequence during fentanyl self-administration, so it is fair to infer that the LHb reports the outcome of a decision made elsewhere. Moreover, since roughly half of the press-responsive neurons at 0.8 mA were also shock-responsive and roughly half of the shock-responsive neurons stayed shock-responsive into Relapse, then a single dominant source would most economically produce this response with the remainder recruited from elsewhere. We suggest that medial prefrontal projections to the LHb are positioned to carry this outcome information and behavioral control (Warden et al., 2012). The remainder could instead be supplied by other afferents such as the 1) Entopeduncular inputs, which are associated with omitted reward and negative prediction error (Stephenson-Jones et al., 2016), 2) Lateral preoptic inputs, which are recruited by noxious stimuli and their predictive cues (Barker et al., 2017), 3) Lateral hypothalamic inputs, which convey homeostatic and motivational state (Stamatakis et al., 2016), and 4) Ventral pallidal inputs, which arrive as opposing glutamatergic and GABAergic populations carrying negative and positive motivational value (Faget et al., 2018;

Stephenson-Jones et al., 2020).

Lastly, as evidenced above, silencing the LHb would test whether downstream readers use its report, and driving it would test whether that report is sufficient to move behavior, but neither would establish whether the report arrives before or after the choice. Furthermore, since the answer is after, a null from either manipulation would have been uninterpretable with respect to the question that was asked, and a positive result would have been read as evidence that the LHb gates a decision that it demonstrably only reports. Therefore the recording did not defer the manipulation so much as determine which manipulation is worth running at all.

5.5 Limitations across the dissertation

MPS focuses on preprocessing and CNMF-based source extraction, so that online decoding is not implemented, distributed cluster-scale processing remains future work, and GPU acceleration is not uniformly available across the submodules. Long recordings with highly dynamic fluorescence or photobleaching may also need baseline correction routines that are not yet in the GUI presets. S2C in turn depends on centroid quality from the upstream pipeline and assumes rigid or similarity inter-session transforms, so that large nonlinear distortions are handled only to the extent that they appear locally rigid within the quad neighborhoods. Its descriptor matching is $O(N^2)$, which is the binding cost at high density, and mesoscope-class recordings of 3,000 to 12,000 neurons sit outside the tested 100 to 1,000 range by design. Since RANSAC fits rigid transforms with scale held fixed, S2C also rejects genuine matches in preparations that expand uniformly between sessions, such as developmental two-photon recordings, and releasing the scale parameter is a one-line change that has not been validated here.

The Ultrasonic Calls cannot be locked to individual trials, since vocalizations are emitted in bouts that are highly likely to reflect a sustained internal state rather than one call per event and the latencies were long and variable relative to a trial. USVs are therefore a session-level and epoch-level measure of affective state, which is the level at which they are used here. The vocal analyses also rest on fewer animals than the behavior, the omnibus test across sessions did not reach significance (Friedman $\chi^2(5) = 10.25$, $p = 0.069$), and the 0.4 mA session contributed a single animal to the peak-timing comparison because vocal suppression at that intensity left no other animal above the call threshold, so the affective results are best read as a pattern that reorganizes with the ladder rather than as a per-animal effect estimated with precision.

Cross-session registration recovers only the subset of neurons visible on multiple days, so that claims about ensemble stability apply to that subset rather than to the structure as a whole. Since GCaMP detection is activity-dependent, a neuron that falls silent in an intervening session cannot be segmented and so drops from the chain, which makes the 41.6% end-to-end DMS retention a conservative lower bound on same-cell persistence rather than a measure of indicator loss. Optogenetic tagging could establish a positive identity for tagged cells, but would also report nothing about false positives on the unlabeled majority. Therefore, we rely on the synthetic benchmark

and the Allen Brain Observatory reproduction for ground-truth accuracy and treat the in-tissue retention as a conservative bound rather than as a per-cell identity we can positively confirm in the recording itself. Moreover, the cross-day ensemble comparisons carry wide confidence intervals, and an overlap that does not exceed chance is not itself evidence that the two labels are independent.

Imaging was restricted to Days 19, 21, 23, and 30, so that the LHb was recorded at 0.2 mA and at 0.8 mA but not at 0.4 mA, which is the highest intensity at which Danger responding was still preserved. That restriction was not a design choice so much as a constraint of running this paradigm for the first time. A faulty component caused intermittent failures of the miniscopes and acquisition hardware over the course of the escalation week, and thus, we recorded sessions of our interest rather than all sessions to preserve the experiment as best as possible. Imaging every rung of the ladder is what a repeat of this experiment should do, and it is the first thing we would change. The ladder was also escalated within animals in a fixed ascending order, so that the effect of a given intensity is not separable from the history of intensities preceding it, and prior exposure to high shock is known to make self-administration more responsive to an intensity that was previously ineffective (Durand et al., 2022). We did not measure flinch or jump thresholds, so we cannot rule out that some of the between-animal variability reflects nociceptive sensitivity rather than a decision process, although shock insensitivity in rats has been traced to a failure of instrumental control rather than to blunted aversion (Jean-Richard-dit Bressel et al., 2019).

5.6 Future directions

Retrograde labeling or projection-specific expression combined with the same imaging would test whether the medial prefrontal cortex supplies the outcome signal that the LHb reports, and the ensemble composition predicts that a single dominant source accounts for roughly half of it. Imaging at 0.4 mA would close the gap between the last intensity at which Danger responding is preserved and the first at which it collapses, and since nothing across 0.1 to 0.4 mA graded with intensity, a finer ladder concentrated at 0.4, 0.5, 0.6, 0.7, and 0.8 mA would localize where the encoding changes far better than the spacing used here. Separating intensity from order would then require running single fixed intensities across repeated days followed by a Relapse, one cohort held at 0.1 mA, another at 0.2, another at 0.4, and another at 0.8, so that the effect of an intensity is read without the history of the intensities below it. Varying the shock so that its probability or magnitude carries prediction error, which is the variable the LHb is classically tuned to, would test whether the retrodictive signal becomes predictive when the outcome is uncertain rather than fixed. Replacing the shock with a social reward would then test whether the same within-session stopping can be driven by an available alternative rather than by an adverse consequence, since social choice procedures already produce voluntary abstinence in drug-free animals (Venniro et al., 2020). Running these recordings in the DMS or the medial prefrontal cortex under the same ladder would show whether the retrodictive dynamics are particular to the LHb or shared by the structures up and downstream of it. Once the upstream source is identified, an optogenetic manipulation

placed there rather than in the LHb would test the decision itself rather than its report, which is the manipulation the retrodictive result actually licenses. For the DMS, pairing S2C tracking with D1 and D2 reporters would in turn resolve whether the cells cycling through the responsive group do so within or across the striatal pathways. Additionally, since the premise of the paradigm is that the decision to stop is made while the drug is on board, then a translation would have to bring a person to that state and then teach a strategy for stopping within it rather than during abstinence, a design outside the expertise of this dissertation and noted here only as the direction the behavioral result suggests.

5.7 Conclusion

The decision to stop depended on a specific consequence being present rather than on stable anti-drug learning, which implies that individuals who achieve cessation through legal sanctions, medical crises, or relationship ultimatums may remain vulnerable in contexts where those consequences seem absent or distant. The adverse consequence itself was not forgotten across the abstinence week, since the shock ensemble at Relapse was recovered at a strength comparable to 0.8 mA, so that treatment does not need to reinstall the knowledge that using carries a cost so much as keep that cost engaged at the moment the choice is actually made. The takeaway is therefore not that an adverse consequence is itself the intervention, since a consequence cannot reasonably be delivered reliably at every point a person decides whether to continue, but that stopping while the drug is on board can be learned at all and that what is learned holds. Reaching this conclusion required a paradigm that could pose the question, tools that could capture it, and a design that manipulated what a clinic can manipulate. Thus, given that these tools are not exclusive to the LHb or addiction, then it is exciting to imagine what will be discovered with them next.

Bibliography

- D. Aharoni and T. M. Hoogland. Circuit investigations with open-source miniaturized microscopes: past, present and future. *Frontiers in Cellular Neuroscience*, 13:141, 2019a. doi: 10.3389/fncel.2019.00141.
- D. Aharoni and T. M. Hoogland. Circuit investigations with open-source miniaturized microscopes: Past, present and future. *Frontiers in Cellular Neuroscience*, 13:141, 2019b. doi: 10.3389/fncel.2019.00141.
- D. Aharoni, B. S. Khakh, A. J. Silva, and P. Golshani. All the light that we can see: A new era in miniaturized microscopy. *Nature Methods*, 16(1):11–13, 2019. doi: 10.1038/s41592-018-0266-x.
- Allen Institute for Brain Science. Allen brain observatory: Visual coding 2p dataset, 2016. URL <https://observatory.brain-map.org/visualcoding>. Accessed via AllenSDK BrainObservatoryCache. [BLANK: access date].
- N. H. Azrin and W. C. Holz. Punishment. In W. K. Honig, editor, *Operant Behavior: Areas of Research and Application*, pages 380–447. Appleton-Century-Crofts, New York, 1966.
- K. R. Bailey and J. N. Crawley. Anxiety-related behaviors in mice. In *Methods of Behavior Analysis in Neuroscience*. CRC Press, 2nd edition, 2009.
- B. W. Balleine, J. Peak, M. Matamalas, J. Bertran-Gonzalez, and G. Hart. The dorsomedial striatum: an optimal cellular environment for encoding and updating goal-directed learning. *Current Opinion in Behavioral Sciences*, 41:38–44, 2021. doi: 10.1016/j.cobeha.2021.03.004.
- G. Barbera, R. Thapa, N. Adhikari, Y. Li, and D. T. Lin. Imaging distinct neuronal populations with a dual-channel miniscope. *Frontiers in Neuroscience*, 18:1445020, 2024. doi: 10.3389/fnins.2024.1445020.
- D. J. Barker, J. Miranda-Barrientos, S. Zhang, D. H. Root, H.-L. Wang, B. Liu, E. S. Calipari, and M. Morales. Lateral preoptic control of the lateral habenula through convergent glutamate and GABA transmission. *Cell Reports*, 21(7):1757–1769, Nov. 2017. doi: 10.1016/j.celrep.2017.10.066.
- N. J. Beacher, M. W. Wang, M. C. Broomer, J. Y. Kuo, K. A. Washington, M. Targum, Y. Zhang, G. Barbera, and D.-T. Lin. Miniscope imaging of nucleus accumbens neural activity in freely

- behaving rats: Virus injection, gradient index lens implantation, recording strategies, and analytical methods. *Current Protocols*, 5(1):e70090, 2025. doi: 10.1002/cpz1.70090.
- B. N. Bender, S. J. Stringfield, and M. M. Torregrossa. Changes in dorsal striatal activity and dopamine release across the development of cocaine seeking habits. *Addiction Neuroscience*, 11: 100149, 2024. doi: 10.1016/j.addicn.2024.100149.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.
- B. S. Bentzley, T. C. Jhou, and G. Aston-Jones. Economic demand predicts addiction-like behavior and therapeutic efficacy of oxytocin in the rat. *Proceedings of the National Academy of Sciences*, 111(32):11822–11827, 2014. doi: 10.1073/pnas.1406324111.
- G. J. Blair, Z. Chen, Z. Dong, C. Guo, J. Zhou, D. Aharoni, Y. Ziv, A. Izquierdo, and H. T. Blair. Hippocampal place cell remapping occurs with memory but not with shock exposure. *eLife*, 12: e80661, 2023. doi: 10.7554/eLife.80661. Rat CA1 imaging with large-field-of-view miniscope; frequently used as a 2023 miniscope benchmark.
- C. Boutsidis and E. Gallopoulos. Svd-based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. doi: 10.1016/j.patcog.2007.09.010.
- BRAIN Initiative. UCLA miniscope — toolmakers. <https://www.braininitiative.org/toolmakers/resources/ucla-miniscope/>, 2025. Accessed 2025.
- R. Brette. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42:e215, 2019. doi: 10.1017/S0140525X19000049.
- S. M. Brudzynski. Ethotransmission: Communication of emotional states through ultrasonic vocalization in rats. *Current Opinion in Neurobiology*, 23(3):310–317, 2013.
- D. J. Cai, D. Aharoni, T. Shuman, J. Shobe, J. Biane, W. Song, and A. J. Silva. A shared neural ensemble links distinct contextual memories encoded close in time. *Nature*, 534:115–118, 2016a. doi: 10.1038/nature17955.
- D. J. Cai, D. Aharoni, T. Shuman, J. Shobe, J. Biane, W. Song, B. Wei, M. Veshkini, M. La-Vu, J. Lou, S. E. Flores, I. Kim, Y. Sano, M. Zhou, K. Baumgaertel, A. Lavi, M. Kamata, M. Tuszyński, M. Mayford, P. Golshani, and A. J. Silva. A shared neural ensemble links distinct contextual memories encoded close in time. *Nature*, 534:115–118, 2016b. doi: 10.1038/nature17955.
- Z. Chen, G. J. Blair, C. Guo, J. Zhou, J. L. Romero-Sosa, A. Izquierdo, and H. T. Blair. A hardware system for real-time decoding of in vivo calcium imaging data. *eLife*, 12:e78344, 2023. doi: 10.7554/eLife.78344.

- J. J. Chow, K. M. Pitts, K. Negishi, R. Madangopal, Y. Dong, M. E. Wolf, and Y. Shaham. Neurobiology of the incubation of drug craving: An update. *Pharmacological Reviews*, 77(2): 100022, 2025. doi: 10.1016/j.pharmrev.2024.100022.
- A. Chu et al. A fear conditioned cue orchestrates a suite of behaviors in rats. *eLife*, 13:e82497, 2024.
- K. R. Coffey, D. J. Barker, S. Ma, and M. O. West. Building an open-source robotic stereotaxic instrument. *Journal of Visualized Experiments*, 80:e51006, 2013. doi: 10.3791/51006.
- K. R. Coffey, R. E. Marx, and J. F. Neumaier. DeepSqueak: A deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859–868, 2019.
- K. R. Coffey, W. B. Nickelson, A. J. Dawkins, and J. F. Neumaier. Rapid appearance of negative emotion during oral fentanyl self-administration in male and female rats. *Addiction Biology*, 28(12):e13344, 2023. doi: 10.1111/adb.13344.
- M. Congiu, M. Trusel, M. Pistis, M. Mameli, and S. Lecca. Opposite responses to aversive stimuli in lateral habenula neurons. *European Journal of Neuroscience*, 50(6):2921–2930, 2019. doi: 10.1111/ejn.14400.
- S. E. J. de Vries, J. A. Lecoq, M. A. Buice, P. A. Groblewski, G. K. Ocker, M. Oliver, D. Feng, N. Cain, P. Ledochowitsch, D. Millman, K. Roll, M. Garrett, T. Keenan, L. Kuan, S. Mihalas, S. Olsen, C. Thompson, W. Wakeman, J. Waters, D. Williams, C. Koch, and R. C. Reid. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020. doi: 10.1038/s41593-019-0550-9.
- V. Deroche-Gamonet, D. Belin, and P. V. Piazza. Evidence for addiction-like behavior in the rat. *Science*, 305(5686):1014–1017, 2004. doi: 10.1126/science.1099020.
- S. Diamond and S. Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- A. V. Domingues, T. T. A. Carvalho, G. J. Martins, R. Correia, B. Coimbra, R. Bastos-Gonçalves, M. Wezik, R. Gaspar, L. Pinto, N. Sousa, R. M. Costa, C. Soares-Cunha, and A. J. Rodrigues. Dynamic representation of appetitive and aversive stimuli in nucleus accumbens shell D1- and D2-medium spiny neurons. *Nature Communications*, 16(1):59, 2025. doi: 10.1038/s41467-024-55269-9.
- Z. Dong, W. Mau, Y. Feng, Z. T. Pennington, L. Chen, Y. Zaki, K. Rajan, T. Shuman, D. Aharoni, and D. J. Cai. Minian, an open-source miniscope analysis pipeline. *eLife*, 11:e70661, 2022a. doi: 10.7554/eLife.70661.
- Z. Dong, W. Mau, Y. Feng, Z. T. Pennington, L. Chen, Y. Zaki, K. Rajan, T. Shuman, D. Aharoni, and D. J. Cai. Minian, an open-source miniscope analysis pipeline. *eLife*, 11:e70661, 2022b. doi: 10.7554/eLife.70661.

- Z. Dong, Y. Feng, K. Diego, A. M. Baggetta, B. M. Sweis, Z. T. Pennington, and D. J. Cai. Simultaneous two-color imaging with a dual-channel miniscope in freely behaving mice. *Science Advances*, 11(27):eadr6470, 2025. doi: 10.1126/sciadv.adr6470.
- L. N. Driscoll, N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170(5):986–999.e16, 2017. doi: 10.1016/j.cell.2017.07.021.
- Y. Duan, P.-J. Tsai, B. J. Salmeron, Y. Hu, H. Gu, H. Lu, J. L. Cadet, E. A. Stein, and Y. Yang. Compulsive drug-taking is associated with habenula-frontal cortex connectivity. *Proceedings of the National Academy of Sciences*, 119(50):e2208867119, 2022. doi: 10.1073/pnas.2208867119.
- A. Durand, P. Girardeau, L. Freese, and S. H. Ahmed. Increased responsiveness to punishment of cocaine self-administration after experience with high punishment. *Neuropsychopharmacology*, 47(2):444–453, 2022. doi: 10.1038/s41386-021-01159-3.
- R. B. Ebitz and B. Y. Hayden. The population doctrine in cognitive neuroscience. *Neuron*, 109(19):3055–3068, 2021. doi: 10.1016/j.neuron.2021.07.011.
- B. Engelhard, J. Finkelstein, J. Cox, W. Fleming, H. J. Jang, S. Ornelas, S. A. Koay, S. Y. Thiberge, N. D. Daw, D. W. Tank, and I. B. Witten. Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, 570(7762):509–513, 2019.
- M. Engeln and S. H. Ahmed. The multiple faces of footshock punishment in animal research on addiction. *Neurobiology of Learning and Memory*, 213:107955, 2024. doi: 10.1016/j.nlm.2024.107955.
- M. Engeln, S. Mitra, R. Chandra, U. Gyawali, M. E. Fox, D. M. Dietz, and M. K. Lobo. Sex-specific role for *Egr3* in nucleus accumbens D2-medium spiny neurons following long-term abstinence from cocaine self-administration. *Biological Psychiatry*, 87(11):992–1000, 2020. doi: 10.1016/j.biopsych.2019.10.019.
- L. Faget, V. Zell, E. Souter, A. McPherson, R. Ressler, N. Gutierrez-Reed, J. H. Yoo, D. Dulcis, and T. S. Hnasko. Opponent control of behavioral reinforcement by inhibitory and excitatory projections from the ventral pallidum. *Nature Communications*, 9(1):849, 2018. doi: 10.1038/s41467-018-03125-y.
- M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. doi: 10.1145/358669.358692.
- J. E. Fragale, C. B. Pantazis, M. H. James, and G. Aston-Jones. The role of orexin-1 receptor signaling in demand for the opioid fentanyl. *Neuropsychopharmacology*, 44(10):1690–1697, 2019. doi: 10.1038/s41386-019-0420-x.

- I. Fredriksson, P.-J. Tsai, A. Shekara, Y. Duan, S. V. Applebey, H. Lu, J. M. Bossert, Y. Shaham, and Y. Yang. Orbitofrontal cortex and dorsal striatum functional connectivity predicts incubation of opioid craving after voluntary abstinence. *Proceedings of the National Academy of Sciences*, 118(43):e2106624118, 2021a. doi: 10.1073/pnas.2106624118.
- I. Fredriksson, M. Venniro, D. J. Reiner, J. J. Chow, J. M. Bossert, and Y. Shaham. Animal models of drug relapse and craving after voluntary abstinence: A review. *Pharmacological Reviews*, 73(3): 1050–1083, 2021b. doi: 10.1124/pharmrev.120.000191.
- I. Fredriksson, M. Venniro, D. J. Reiner, J. J. Chow, J. M. Bossert, and Y. Shaham. Animal models of drug relapse and craving after voluntary abstinence: A review. *Pharmacological Reviews*, 73(3): 1050–1083, 2021c. doi: 10.1124/pharmrev.120.000191.
- J. Freeman. Open source tools for large-scale neuroscience. *Nature Neuroscience*, 18(10):1321–1332, 2015. doi: 10.1038/nn.4088.
- M. F. Garnett and A. M. Miniño. Drug overdose deaths in the United States, 2003–2023. NCHS Data Brief 522, National Center for Health Statistics, 2024.
- C. R. Gerfen and D. J. Surmeier. Modulation of striatal projection systems by dopamine. *Annual Review of Neuroscience*, 34:441–466, 2011. doi: 10.1146/annurev-neuro-061010-113641.
- K. K. Ghosh, L. D. Burns, E. D. Cocker, A. Nimmerjahn, Y. Ziv, A. El Gamal, and M. J. Schnitzer. Miniaturized integration of a fluorescence microscope. *Nature Methods*, 8:871–878, 2011a. doi: 10.1038/nmeth.1694.
- K. K. Ghosh, L. D. Burns, E. D. Cocker, A. Nimmerjahn, Y. Ziv, A. El Gamal, and M. J. Schnitzer. Miniaturized integration of a fluorescence microscope. *Nature Methods*, 8:871–878, 2011b. doi: 10.1038/nmeth.1694.
- A. Giovannucci, J. Friedrich, P. Gunn, J. Kalfon, B. L. Brown, S. A. Koay, and E. A. Pnevmatikakis. Caiman: An open source tool for scalable calcium imaging data analysis. *eLife*, 8:e38173, 2019a. doi: 10.7554/eLife.38173.
- A. Giovannucci, J. Friedrich, P. Gunn, J. Kalfon, B. L. Brown, S. A. Koay, J. Taxidis, F. Najafi, J. L. Gauthier, P. Zhou, B. S. Khakh, D. W. Tank, D. B. Chklovskii, and E. A. Pnevmatikakis. CaImAn: an open source tool for scalable calcium imaging data analysis. *eLife*, 8:e38173, 2019b. doi: 10.7554/eLife.38173.
- J. Goffinet, S. Brudner, R. Mooney, and J. Pearson. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*, 10:e67855, 2021.
- B. O. Gomez, E. Y. Samayoa, P. O. Coffin, L. W. Suen, S. L. Ivey, K. R. Knight, C. McCuistian, A. Appa, and A. R. Bazazi. Mechanisms of overdose from unintentional fentanyl use when

- intending to use stimulants: a qualitative study of people who use stimulants and avoid opioids. *Research Square*, Feb. 2026. doi: 10.21203/rs.3.rs-8555473/v1. Preprint.
- W. G. Gonzalez, H. Zhang, A. Harutyunyan, and C. Lois. Persistence of hippocampal neural representations through time and damage. *Science*, 365(6455):821–825, 2019. doi: 10.1126/science.aat6925.
- N. M. Graziane, S. Sun, W. J. Wright, D. Jang, Z. Liu, Y. H. Huang, E. J. Nestler, Y. T. Wang, O. M. Schlüter, and Y. Dong. Opposing mechanisms mediate morphine- and cocaine-induced generation of silent synapses. *Nature Neuroscience*, 19(7):915–925, 2016. doi: 10.1038/nn.4313.
- S. Gu, E. L. Mackevicius, M. S. Fee, and P. Zhou. Spatial tracking across time (STAT): Tracking neurons across in-vivo imaging sessions through optimizing local neighborhood motion consistency. *bioRxiv*, 2023. doi: 10.1101/2023.05.13.540658.
- M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup. Efficient subpixel image registration algorithms. *Optics Letters*, 33(2):156–158, 2008. doi: 10.1364/OL.33.000156.
- C. Guo, A. Das, J. Provoste-Olivares, et al. Miniscope-lfov: A large-field-of-view, single-cell resolution, miniature microscope for population imaging in freely behaving animals. *Science Advances*, 9(44):ædg3918, 2023. doi: 10.1126/sciadv.adg3918.
- N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.
- X. Han, L. Albrechet-Souza, M. R. Doyle, A. Shimamoto, J. F. DeBold, and K. A. Miczek. Social stress and escalated drug self-administration in mice II. Cocaine and dopamine in the nucleus accumbens. *Psychopharmacology*, 232(6):1003–1010, 2015.
- K. D. Harris. A shift test for independence in generic time series, 2020.
- Y. Hashikawa, K. Hashikawa, M. A. Rossi, M. L. Basiri, Y. Liu, N. L. Johnston, O. R. Ahmad, and G. D. Stuber. Transcriptional and spatial resolution of cell types in the mammalian habenula. *Neuron*, 106(5):743–758.e5, 2020.
- R. Hattori and T. Komiyama. PatchWarp: Corrections of non-uniform image distortions in two-photon calcium imaging data by patchwork affine transformations. *Cell Reports Methods*, 2(5):100205, 2022. doi: 10.1016/j.crmeth.2022.100205.
- M. C. Hearing, J. Jedynak, S. R. Ebner, A. Ingebretson, A. J. Asp, R. A. Fischer, C. Schmidt, E. B. Larson, and M. J. Thomas. Reversal of morphine-induced cell-type-specific synaptic plasticity in the nucleus accumbens shell blocks reinstatement. *Proceedings of the National Academy of Sciences*, 113(3):757–762, 2016. doi: 10.1073/pnas.1519248113.

- O. Hikosaka. The habenula: From stress evasion to value-based decision-making. *Nature Reviews Neuroscience*, 11(7):503–513, 2010. doi: 10.1038/nrn2866.
- O. Hikosaka, S. R. Sesack, L. Lecourtier, and P. D. Shepard. Habenula: Crossroad between the basal ganglia and the limbic system. *Journal of Neuroscience*, 28(46):11825–11829, 2008.
- D. V. Hinkley and E. A. Hinkley. Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488, 1970. doi: 10.1093/biomet/57.3.477.
- D. W. Hogg, M. Blanton, D. Lang, K. Mierle, and S. Roweis. Automated astrometry. In R. W. Argyle, P. S. Bunclark, and J. R. Lewis, editors, *Astronomical Data Analysis Software and Systems XVII*, volume 394 of *ASP Conference Series*, pages 27–34, 2008.
- K. S. Holly, C. O. Orndorff, and T. A. Murray. MATSAP: An automated analysis of stretch-attend posture in rodent behavioral experiments. *Scientific Reports*, 6:31286, 2016.
- A. Holtmaat, T. Bonhoeffer, D. K. Chow, J. Chuckowree, V. De Paola, S. B. Hofer, M. Hübener, T. Keck, G. Knott, W.-C. A. Lee, R. Mostany, T. D. Mrsic-Flogel, E. Nedivi, C. Portera-Cailliau, K. Svoboda, J. T. Trachtenberg, and L. Wilbrecht. Long-term, high-resolution imaging in the mouse neocortex through a chronic cranial window. *Nature Protocols*, 4(8):1128–1144, 2009. doi: 10.1038/nprot.2009.89.
- S. Hong, T. C. Jhou, M. Smith, K. S. Saleem, and O. Hikosaka. Negative reward signals from the lateral habenula to dopamine neurons are mediated by rostromedial tegmental nucleus in primates. *Journal of Neuroscience*, 31(32):11457–11471, 2011. doi: 10.1523/JNEUROSCI.1384-11.2011.
- H. Hu, Y. Cui, and Y. Yang. Circuits and functions of the lateral habenula in health and in disease. *Nature Reviews Neuroscience*, 21(5):277–295, 2020. doi: 10.1038/s41583-020-0292-4.
- S. R. Hursh and P. G. Roma. Behavioral economics and empirical public policy. *Journal of the Experimental Analysis of Behavior*, 99(1):98–124, 2013. doi: 10.1002/jeab.7.
- S. R. Hursh and A. Silberberg. Economic demand and essential value. *Psychological Review*, 115(1):186–198, 2008. doi: 10.1037/0033-295X.115.1.186.
- H. Inan, C. Schmuckermair, T. Tasci, B. O. Ahanonu, O. Hernandez, J. Lecoq, F. Dinç, M. J. Wagner, M. A. Erdogdu, and M. J. Schnitzer. Fast and statistically robust cell extraction from large-scale neural calcium imaging datasets. *bioRxiv*, 2021a. doi: 10.1101/2021.03.24.436279.
- H. Inan, C. Schmuckermair, T. Tasci, B. O. Ahanonu, O. Hernandez, J. Lecoq, F. Dinç, M. J. Wagner, M. A. Erdogdu, and M. J. Schnitzer. Fast and statistically robust cell extraction from large-scale neural calcium imaging datasets. *bioRxiv*, 2021b. doi: 10.1101/2021.03.24.436279.
- P. Jean-Richard-dit Bressel and G. P. McNally. The role of the lateral habenula in punishment. *PLoS ONE*, 9(11):e111699, 2014. doi: 10.1371/journal.pone.0111699.

- P. Jean-Richard-dit Bressel, C. Ma, L. A. Bradfield, S. Killcross, and G. P. McNally. Punishment insensitivity emerges from impaired contingency detection, not aversion insensitivity or reward dominance. *eLife*, 8:e52765, 2019. doi: 10.7554/eLife.52765.
- J. H. Jennings, R. L. Ung, S. L. Resendez, A. M. Stamatakis, J. G. Taylor, J. Huang, K. Veleta, P. A. Kantak, M. Aita, K. Shilling-Scriver, C. Ramakrishnan, K. Deisseroth, S. Otte, and G. D. Stuber. Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. *Nature*, 565(7741):645–649, 2019. doi: 10.1038/s41586-019-1493-4.
- T. C. Jhou, H. L. Fields, M. G. Baxter, C. B. Saper, and P. C. Holland. The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. *Neuron*, 61(5):786–800, 2009. doi: 10.1016/j.neuron.2009.02.001.
- K. G. Johnston, S. F. Grieco, H. Zhang, S. Jin, X. Xu, and Q. Nie. Tracking longitudinal population dynamics of single neuronal calcium signal using SCOUT. *Cell Reports Methods*, 2(5):100207, 2022. doi: 10.1016/j.crmeth.2022.100207.
- S. Jonkman, Y. Pelloux, and B. J. Everitt. Differential roles of the dorsolateral and dorsomedial striatum in punished cocaine seeking. *Journal of Neuroscience*, 32(13):4645–4650, 2012. doi: 10.1523/JNEUROSCI.6661-11.2012.
- M. N. Koffarnus, C. T. Franck, J. S. Stein, and W. K. Bickel. A modified exponential behavioral economic demand model to better describe consumption data. *Experimental and Clinical Psychopharmacology*, 23(6):504–512, 2015. doi: 10.1037/pha0000045.
- M.-S. Kong, E. Ancell, D. M. Witten, and L. S. Zweifel. Valence and salience encoding in the central amygdala. *eLife*, 13:RP101980, 2025. doi: 10.7554/eLife.101980.
- G. F. Koob. Drug addiction: Hyperkatifeia/negative reinforcement as a framework for medications development. *Pharmacological Reviews*, 73(1):163–201, 2021.
- J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017. doi: 10.1016/j.neuron.2016.12.041.
- A. C. Kreitzer and R. C. Malenka. Striatal plasticity and basal ganglia circuit function. *Neuron*, 60(4):543–554, 2008. doi: 10.1016/j.neuron.2008.11.005.
- N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540, 2009. doi: 10.1038/nm.2303.
- H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

- T. Lagache, A. Hanson, J. E. Pérez-Ortega, A. Fairhall, and R. Yuste. Tracking calcium dynamics from individual neurons in behaving animals. *PLoS Computational Biology*, 17(10):e1009432, 2021. doi: 10.1371/journal.pcbi.1009432.
- D. Lang, D. W. Hogg, K. Mierle, M. Blanton, and S. Roweis. Astrometry.net: blind astrometric calibration of arbitrary astronomical images. *The Astronomical Journal*, 139:1782–1800, 2010. doi: 10.1088/0004-6256/139/5/1782.
- S. Lecca, F. J. Meye, M. Trusel, A. Tchenio, J. Harris, M. K. Schwarz, D. Burdakov, F. Georges, and M. Mameli. Aversive stimuli drive hypothalamus-to-habenula excitation to promote escape behavior. *eLife*, 6:e30697, 2017. doi: 10.7554/eLife.30697.
- A. A. Legaria, B. A. Matikainen-Ankney, B. Yang, B. Ahanonu, J. A. Licholai, J. G. Parker, and A. V. Kravitz. Fiber photometry in striatum reflects primarily nonsomatic changes in calcium. *Nature Neuroscience*, 25(9):1124–1128, 2022. doi: 10.1038/s41593-022-01152-z.
- F. J. Leong, M. Brady, and J. O. McGee. Correction of uneven illumination (vignetting) in digital microscopy images. *Journal of Clinical Pathology*, 56(8):619–621, 2003. doi: 10.1136/jcp.56.8.619.
- J. P. Lewis. Fast normalized cross-correlation. In *Vision Interface*, pages 120–123, 1995.
- B. Li, J. Piriz, M. Mirrione, C. Chung, C. D. Proulx, D. Schulz, F. Henn, and R. Malinow. Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature*, 470(7335):535–539, 2011.
- C. Li, X. Yang, Y. Ke, and W.-H. Yung. Fully affine invariant methods for cross-session registration of calcium imaging data. *eNeuro*, 7(4):ENEURO.0054–20.2020, 2020. doi: 10.1523/ENEURO.0054-20.2020.
- LifeCanvas Technologies. Smartbatch+ protocol outline. <https://sites.google.com/lifecanvastech.com/protocol/outline>, n.d. Accessed 2025-12-22.
- J. Lu, C. Li, J. Singh-Alvarado, Z. C. Zhou, F. Fröhlich, R. Mooney, and F. Wang. Min1pipe: A miniscope 1-photon-based calcium imaging signal extraction pipeline. *Cell Reports*, 23(12):3673–3684, 2018. doi: 10.1016/j.celrep.2018.05.062.
- Z. Ma, Y. Duan, I. Fredriksson, P.-J. Tsai, A. Batista, H. Lu, Y. Shaham, and Y. Yang. Role of dorsal striatum circuits in relapse to opioid seeking after voluntary abstinence. *Neuropsychopharmacology*, 50(2):452–460, 2025. doi: 10.1038/s41386-024-01990-4.
- J. Majnik, M. Mantez, S. Zangila, S. Bugeon, L. Guignard, J.-C. Platel, and R. Cossart. Longitudinal tracking of neuronal activity from the same cells in the developing brain using Track2p. *eLife*, 14:RP107540, 2025. doi: 10.7554/eLife.107540.

- N. J. Marchant, T. N. Khuc, C. L. Pickens, A. Bonci, and Y. Shaham. Context-induced relapse after extinction versus punishment: similarities and differences. *Psychopharmacology*, 226(3): 485–498, 2013. doi: 10.1007/s00213-012-2913-7.
- N. J. Marchant, E. J. Campbell, and K. Kaganovsky. Punishment of alcohol-reinforced responding in alcohol preferring P rats reveals a bimodal population: implications for models of compulsive drug seeking. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 87(Pt A):68–77, 2018. doi: 10.1016/j.pnpbp.2017.07.020.
- M. Matsumoto and O. Hikosaka. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, 447(7148):1111–1115, 2007.
- M. Matsumoto and O. Hikosaka. Representation of negative motivational value in the primate lateral habenula. *Nature Neuroscience*, 12(1):77–84, 2009.
- S. A. McConnell, A. J. Brandner, B. A. Blank, D. N. Kearns, G. F. Koob, L. F. Vendruscolo, and B. J. Tunstall. Demand for fentanyl becomes inelastic following extended access to fentanyl vapor self-administration. *Neuropharmacology*, 182:108355, 2021. doi: 10.1016/j.neuropharm.2020.108355.
- L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.
- G. P. McNally, M. L. Spetch, J. J. Chow, L. Hogarth, and N. J. Marchant. Pathways to the persistence of drug use despite its adverse consequences. *Molecular Psychiatry*, 28(6):2228–2237, 2023.
- K. A. Miczek and N. H. Mutschler. Activational effects of social stress on IV cocaine self-administration in rats. *Psychopharmacology*, 128(3):256–264, 1996.
- A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. doi: 10.1109/TPAMI.2010.46.
- M. Pachitariu, C. Stringer, M. Dipoppa, S. Schröder, L. F. Rossi, H. Dagleish, M. Carandini, and K. D. Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv*, page 061507, 2017. doi: 10.1101/061507.
- H.-T. Pan, S. Menacherry, and J. Justice, Joseph B. Differences in the pharmacokinetics of cocaine in naive and cocaine-experienced rats. *Journal of Neurochemistry*, 56(4):1299–1306, 1991. doi: 10.1111/j.1471-4159.1991.tb11425.x.
- Y.-G. Park, C. H. Sohn, R. Chen, M. McCue, D. H. Yun, G. T. Drummond, T. Ku, N. B. Evans, H. C. Oak, W. Trieu, H. Choi, X. Jin, V. Lilascharoen, J. Wang, M. C. Truttmann, H. W. Qi, H. L. Ploegh, T. R. Golub, S.-C. Chen, M. P. Frosch, H. J. Kulik, B. K. Lim, and K. Chung. Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nature Biotechnology*, 2018. doi: 10.1038/nbt.4281. Online ahead of print.

- V. Pascoli, A. Hiver, Y. Li, M. Harada, V. Esmaeili, and C. Lüscher. Cell-type specific synaptic plasticity in dorsal striatum is associated with punishment-resistance compulsive-like cocaine self-administration in mice. *Neuropsychopharmacology*, 48(3):448–458, 2023. doi: 10.1038/s41386-022-01429-8.
- G. Paxinos and C. Watson. *The rat brain in stereotaxic coordinates*. Academic Press, 7th edition, 2014.
- A. Peden-Asarch, L. Honan, J. Bai, E. Asarch, J. Quinn, K. R. Coffey, and J. F. Neumaier. Analysis code for: Stars2cells – astrometric tracking of neurons across imaging sessions. Zenodo, v1.0.0, 2026a. URL <https://doi.org/10.5281/zenodo.20938869>.
- A. Peden-Asarch, L. Honan, J. Bai, E. Asarch, J. Quinn, K. R. Coffey, and J. F. Neumaier. Data for: Stars2cells – astrometric tracking of neurons across imaging sessions. Zenodo, v1.0.0, 2026b. URL <https://doi.org/10.5281/zenodo.20938743>.
- A. Peden-Asarch, M. Weinstock, K. R. Coffey, and J. F. Neumaier. Miniscope processing suite (MPS): an intuitive, no-code, scalable pipeline for long-duration calcium imaging. *eLife*, 15: RP109858, 2026c. doi: 10.7554/eLife.109858.1.
- A. M. Peden-Asarch, M. Weinstock, K. R. Coffey, and J. F. Neumaier. Miniscope processing suite (MPS): An intuitive, no-code, scalable pipeline for long duration calcium imaging. *bioRxiv*, 2025. doi: 10.64898/2025.12.01.691517.
- A. M. Peden-Asarch, L. E. Honan, J. Z. Bai, E. M. Asarch, J. A. Quinn, K. R. Coffey, and J. F. Neumaier. Stars2Cells: Astrometric tracking of neurons across imaging sessions. *bioRxiv*, 2026d. doi: 10.64898/2026.07.03.736144.
- A. M. Peden-Asarch, L. E. Honan, J. Z. Bai, E. M. Asarch, J. A. Quinn, K. R. Coffey, and J. F. Neumaier. Stars2Cells: Astrometric tracking of neurons across imaging sessions. *bioRxiv*, 2026e. doi: 10.64898/2026.07.03.736144.
- Y. Pelloux, B. J. Everitt, and A. Dickinson. Compulsive drug seeking by rats under punishment: effects of drug taking history. *Psychopharmacology*, 194(1):127–137, 2007. doi: 10.1007/s00213-007-0805-0.
- Photonics Media. Bruker acquires Inscopix. <https://www.photonics.com/Articles/a68503>, 2022. Accessed 2025.
- P. V. Piazza and M. Le Moal. The role of stress in drug self-administration. *Trends in Pharmacological Sciences*, 19(2):67–74, 1998.
- C. L. Pickens, M. Airavaara, F. Theberge, S. Fanous, B. T. Hope, and Y. Shaham. Neurobiology of the incubation of drug craving. *Trends in Neurosciences*, 34(8):411–420, 2011.

- E. A. Pnevmatikakis and A. Giovannucci. Normcorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *Journal of Neuroscience Methods*, 291:83–94, 2017. doi: 10.1016/j.jneumeth.2017.07.031.
- E. A. Pnevmatikakis, Y. Ahmadian, B. Percival, and L. Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016a. doi: 10.1016/j.neuron.2015.11.037.
- E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, M. Ahrens, R. Bruno, T. M. Jessell, D. S. Peterka, R. Yuste, and L. Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016b. doi: 10.1016/j.neuron.2015.11.037.
- C. D. Proulx, O. Hikosaka, and R. Malinow. Reward processing by the lateral habenula in normal and depressive behaviors. *Nature Neuroscience*, 17(9):1146–1152, 2014.
- N. Renier, Z. Wu, D. J. Simon, J. Yang, P. Ariel, and M. Tessier-Lavigne. iDISCO: A simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell*, 159(4):896–910, 2014.
- J. B. Ritchie, D. M. Kaplan, and C. Klein. Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607, 2019. doi: 10.1093/bjps/axx023.
- M. E. Rule, T. O’Leary, and C. D. Harvey. Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147, 2019. doi: 10.1016/j.conb.2019.08.005.
- C. E. Schoonover, S. N. Ohashi, R. Axel, and A. J. P. Fink. Representational drift in primary olfactory cortex. *Nature*, 594:541–546, 2021. doi: 10.1038/s41586-021-03628-7.
- J. L. Seiler, C. V. Cosme, V. N. Sherathiya, M. D. Schaid, J. M. Bianco, A. S. Bridgemohan, and T. N. Lerner. Dopamine signaling in the dorsomedial striatum promotes compulsive behavior. *Current Biology*, 32(5):1175–1188.e5, 2022. doi: 10.1016/j.cub.2022.01.055.
- S. J. Shabel, C. Wang, B. Monk, S. Aronson, and R. Malinow. Stress transforms lateral habenula reward responses into punishment signals. *Proceedings of the National Academy of Sciences*, 116(25):12488–12493, 2019. doi: 10.1073/pnas.1903334116.
- Y. Shaham and J. Stewart. Exposure to mild stress enhances the reinforcing efficacy of intravenous heroin self-administration in rats. *Psychopharmacology*, 114(3):523–527, 1994.
- U. Shalev, J. W. Grimm, and Y. Shaham. Neurobiology of relapse to heroin and cocaine seeking: A review. *Pharmacological Reviews*, 54(1):1–42, 2002.
- L. Sheintuch, A. Rubin, N. Brande-Eilat, N. Geva, N. Sadeh, B. Pinchasof, and Y. Ziv. Tracking the same neurons across multiple days in Ca^{2+} imaging data. *Cell Reports*, 21(4):1102–1115, 2017a. doi: 10.1016/j.celrep.2017.10.013.

- L. Sheintuch, A. Rubin, N. Brande-Eilat, N. Geva, N. Sadeh, O. Pinchasof, and Y. Ziv. Tracking the same neurons across multiple days in Ca^{2+} imaging data. *Cell Reports*, 21:1102–1115, 2017b. doi: 10.1016/j.celrep.2017.10.013.
- G. G. Simpson. Notes on the measurement of faunal resemblance. *American Journal of Science*, 258(2):300–311, 1960.
- N. J. Sofroniew, D. Flickinger, J. King, and K. Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife*, 5:e14472, 2016. doi: 10.7554/eLife.14472.
- A. M. Stamatakis, M. Van Swieten, M. L. Basiri, G. A. Blair, P. Kantak, and G. D. Stuber. Lateral hypothalamic area glutamatergic neurons and their projections to the lateral habenula regulate feeding and reward. *Journal of Neuroscience*, 36(2):302–311, 2016.
- M. Stephenson-Jones, K. Yu, S. Ahrens, J. M. Tucciarone, A. N. van Huijstee, L. A. Mejia, M. A. Penzo, L.-H. Tai, L. Wilbrecht, and B. Li. A basal ganglia circuit for evaluating action outcomes. *Nature*, 539(7628):289–293, 2016. doi: 10.1038/nature19845.
- M. Stephenson-Jones, C. Bravo-Rivera, S. Ahrens, A. Furlan, X. Xiao, C. Fernandes-Henriques, and B. Li. Opposing contributions of GABAergic and glutamatergic ventral pallidal neurons to motivational behaviors. *Neuron*, 105(5):921–933, 2020. doi: 10.1016/j.neuron.2019.12.006.
- S. R. Sternberg. Biomedical image processing. *Computer*, 16(1):22–34, 1983. doi: 10.1109/MC.1983.1654163.
- J. N. Stirman, I. T. Smith, M. W. Kudenov, and S. L. Smith. Wide field-of-view, multi-region, two-photon imaging of neuronal activity in the mammalian brain. *Nature Biotechnology*, 34: 857–862, 2016. doi: 10.1038/nbt.3594.
- C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019. doi: 10.1126/science.aav7893.
- P. Thévenaz, U. E. Ruttimann, and M. Unser. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing*, 7(1):27–41, 1998. doi: 10.1109/83.650848.
- F. Tian, B. Mattison, and W. Yang. Deepinminiscope: Deep learning-powered physics-informed integrated miniscope. *Science Advances*, 11(37):eadr6687, 2025. doi: 10.1126/sciadv.adr6687.
- L. Tiroshi, Y. Atamna, N. Gilin, N. Berkowitz, and J. A. Goldberg. Striatal neurons are recruited dynamically into collective representations of self-initiated and learned actions in freely moving mice. *eNeuro*, 11(1):ENEURO.0315–23.2023, 2024. doi: 10.1523/ENEURO.0315-23.2023.
- J. T. Trachtenberg, B. E. Chen, G. W. Knott, G. Feng, J. R. Sanes, E. Welker, and K. Svoboda. Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. *Nature*, 420:788–794, 2002. doi: 10.1038/nature01273.

- S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: Image processing in Python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.
- M. Venniro, D. Caprioli, and Y. Shaham. Animal models of drug relapse and craving: From drug priming-induced reinstatement to incubation of craving after voluntary abstinence. In *Progress in Brain Research*, volume 224, pages 25–52. Elsevier, 2016.
- M. Venniro, M. L. Banks, M. Heilig, D. H. Epstein, and Y. Shaham. Improving translation of animal models of addiction and relapse by reverse translation. *Nature Reviews Neuroscience*, 21(11):625–643, 2020. doi: 10.1038/s41583-020-0378-z.
- P. Vergara, Y. Wang, S. Srinivasan, Z. Dong, Y. Feng, I. Koyanagi, D. Kumar, Y. Chérasse, T. Naoi, Y. Sugaya, T. Sakurai, M. Kano, T. Shuman, D. Cai, M. Yanagisawa, and M. Sakaguchi. A comprehensive suite for extracting neuron signals across multiple sessions in one-photon calcium imaging. *Nature Communications*, 16(1):3443, 2025a. doi: 10.1038/s41467-025-58817-z.
- P. Vergara, Y. Wang, S. Srinivasan, et al. A comprehensive suite for extracting neuron signals across multiple sessions in one-photon calcium imaging. *Nature Communications*, 16:3443, 2025b. doi: 10.1038/s41467-025-58817-z.
- L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991. doi: 10.1109/34.87344.
- N. D. Volkow and C. Blanco. The changing opioid crisis: development, challenges and opportunities. *Molecular Psychiatry*, 26(1):218–233, 2021. doi: 10.1038/s41380-020-0661-4.
- B. Vollmayr and F. A. Henn. Learned helplessness in the rat: improvements in validity and reliability. *Brain Research Protocols*, 8(1):1–7, 2001. doi: 10.1016/S1385-299X(01)00067-8.
- M. L. Wallace, A. Saunders, K. W. Huang, A. C. Philson, M. Goldman, E. Z. Macosko, S. A. McCarroll, and B. L. Sabatini. Anatomical and single-cell transcriptional profiling of the murine habenular complex. *eLife*, 9:e51271, 2020.
- D. Wang, Y. Li, Q. Feng, Q. Guo, J. Zhou, and M. Luo. Learning shapes the aversion and reward responses of lateral habenula neurons. *eLife*, 6:e23045, 2017.
- M. R. Warden, A. Selimbeyoglu, J. J. Mirzabekov, M. Lo, K. R. Thompson, S.-Y. Kim, A. Adhikari, K. M. Tye, L. M. Frank, and K. Deisseroth. A prefrontal cortex-brainstem neuronal projection that controls response to behavioural challenge. *Nature*, 492(7429):428–432, 2012. doi: 10.1038/nature11617.
- B. R. White, C. M. Lee, and A. Z. Snyder. An open-source framework for calcium imaging data analysis. *Frontiers in Cellular Neuroscience*, 14:588, 2020. doi: 10.3389/fncel.2020.00588.

- H. S. Wirtshafter and J. F. Disterhoft. In vivo multi-day calcium imaging of ca1 hippocampus in freely moving rats reveals a high preponderance of place cells with consistent place fields. *Journal of Neuroscience*, 42(22):4538–4554, 2022. doi: 10.1523/JNEUROSCI.1750-21.2022. Often cited in miniscope literature for long-duration one-photon CA1 imaging in rats.
- Y. Yang, Y. Cui, K. Sang, Y. Dong, Z. Ni, S. Ma, and H. Hu. Ketamine blocks bursting in the lateral habenula to rapidly relieve depression. *Nature*, 554(7692):317–322, 2018. doi: 10.1038/nature25509.
- H. H. Yin and B. J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476, 2006. doi: 10.1038/nrn1919.
- S. Yun and J. G. Parker. Cellular-resolution and bulk-fluorescence recordings of calcium activity yield reciprocal readouts of in vivo drug efficacy. *Synapse*, 79(2):e70011, 2025.
- J. E. Zachry, M. G. Kutlu, H. J. Yoon, M. Z. Leonard, M. Chevé, D. D. Patel, A. Gaidici, V. Kondev, K. C. Thibeault, R. Bethi, J. Tat, P. R. Melugin, A. U. Isiktas, M. E. Joffe, D. J. Cai, P. J. Conn, B. A. Grueter, and E. S. Calipari. D1 and d2 medium spiny neurons in the nucleus accumbens core have distinct and valence-independent roles in learning. *Neuron*, 112(5):835–849.e7, 2024. doi: 10.1016/j.neuron.2023.11.023.
- Y. Zhang, M. Rózsa, Y. Liang, D. Bushey, Z. Wei, J. Zheng, D. Reep, G. J. Broussard, A. Tsang, G. Tsegaye, S. Narayan, C. J. Obara, J.-X. Lim, R. Patel, R. Zhang, M. B. Ahrens, G. C. Turner, S. S.-H. Wang, W. L. Korff, E. R. Schreiter, K. Svoboda, J. P. Hasseman, I. Kolb, and L. L. Looger. Fast and sensitive GCaMP calcium indicators for imaging neural populations. *Nature*, 615(7954):884–891, 2023a. doi: 10.1038/s41586-023-05828-9.
- Y. Zhang, M. Rózsa, Y. Liang, D. Bushey, Z. Wei, J. Zheng, D. Reep, G. J. Broussard, A. Tsang, G. Tsegaye, et al. Fast and sensitive GCaMP calcium indicators for imaging neural populations. *Nature*, 615(7954):884–891, 2023b. doi: 10.1038/s41586-023-05828-9. URL <https://www.nature.com/articles/s41586-023-05828-9>.
- P. Zhou, S. L. Resendez, J. Rodriguez-Romaguera, J. C. Jimenez, S. Q. Neufeld, A. Giovannucci, J. Friedrich, E. A. Pnevmatikakis, G. D. Stuber, R. Hen, M. A. Kheirbek, B. L. Sabatini, R. E. Kass, and L. Paninski. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife*, 7:e28728, 2018a. doi: 10.7554/eLife.28728.
- P. Zhou, S. L. Resendez, J. Rodriguez-Romaguera, J. C. Jimenez, S. Q. Neufeld, A. Giovannucci, and L. Paninski. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife*, 7:e28728, 2018b. doi: 10.7554/eLife.28728.
- P. Zhou, J. H. Macke, S. Turaga, and L. Paninski. EASE: EM-assisted source extraction from calcium imaging data. *bioRxiv*, page 2020.03.25.007468, 2020. doi: 10.1101/2020.03.25.007468. Preprint.

- Y. Ziv and K. K. Ghosh. Miniature microscopes for large-scale imaging of neuronal activity in freely behaving rodents. *Current Opinion in Neurobiology*, 32:141–147, 2015a. doi: 10.1016/j.conb.2015.04.001.
- Y. Ziv and K. K. Ghosh. Miniature microscopes for large-scale imaging of neuronal activity in freely behaving rodents. *Current Opinion in Neurobiology*, 32:141–147, 2015b. doi: 10.1016/j.conb.2015.04.001.
- Y. Ziv, L. D. Burns, E. D. Cocker, E. O. Hamel, K. K. Ghosh, L. J. Kitch, A. El Gamal, and M. J. Schnitzer. Long-term dynamics of CA1 hippocampal place codes. *Nature Neuroscience*, 16: 264–266, 2013a. doi: 10.1038/nn.3329.
- Y. Ziv, L. D. Burns, E. D. Cocker, E. O. Hamel, K. K. Ghosh, L. J. Kitch, A. El Gamal, and M. J. Schnitzer. Long-term dynamics of ca1 hippocampal place codes. *Nature Neuroscience*, 16(3): 264–266, 2013b. doi: 10.1038/nn.3329.