Inferring brain-wide interactions using data-constrained recurrent neural network models

Matthew G. Perich[1,2,*], Charlotte Arlt[3], Sofia Soares[3], Siyan Zhou[3], Manuel Beiran[4], Aaron S. Andalman[5], Tyler Benster[6], Megan E. Young[7], Clayton P. Mosher[7,8], Juri Minxha[8,9], Eugene Carter[3], Ueli Rutishauser[8,9], Peter H. Rudebeck[7], Christopher D. Harvey[3], Karl Deisseroth[5,11,12,13,*], and Kanaka Rajan[3,10,13,*]

[1]Département des neurosciences, Université de Montréal, Montréal, QC, Canada
[2]Quebec Artificial Intelligence Institute (Mila), Montréal, QC, Canada
[3]Department of Neurobiology, Harvard Medical School, Boston, MA, USA
[4]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA
[5]Department of Bioengineering, Stanford University, Stanford, CA, USA
[6]Neurosciences Graduate Program, Stanford University, Stanford, CA, USA
[7]Icahn School of Medicine at Mount Sinai, New York, NY, USA
[8]Cedars-Sinai Medical Center, Los Angeles, CA, USA
[9]California Institute of Technology, Pasadena, CA, USA
[10]Kempner Institute, Boston, MA, USA
[11]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
[12]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA
[13]These authors jointly supervised this work

Behavior arises from the coordinated activity across a number of anatomically and functionally distinct brain regions[1,2]. Modern experimental tools[3–5] allow unprecedented access to large neural populations, even those spanning many interacting regions brain-wide[2]. Yet, understanding such large-scale datasets necessitates both robust, scalable computational models to extract meaningful features of inter-region communication, as well as principled theories to interpret those features. Here, we introduce Current-Based Decomposition (CURBD), an approach for inferring brain-wide interactions using data-constrained recurrent neural network models[6] that, once trained, autonomously produce dynamics consistent with experimentally-obtained neural data. CURBD leverages the functional interactions inferred from such models to reveal directional currents between multiple brain regions simultaneously. We first show that CURBD accurately isolates inter-region currents in simulated, ground-truth networks with known connectivity and dynamics. We then apply CURBD to multi-region neural recordings obtained from a broad range of neural datasets—larval zebrafish[7], mice[8], macaques[9], and humans[10]—to demonstrate the widespread applicability of CURBD in untangling brain-wide interactions and inter-area communication principles underlying behavior.

During development, the nervous systems of even small organisms organize into remarkably complex structures. Brains can exhibit structural modularity (e.g., regions, layers, cell types) with phylogenetically-determined specialization across modules,[11] or alternatively functional modularity with individual brain regions having interdigitated submodules based on the response properties of the component neurons. Brain regions, even if we define them purely anatomically, have striking specialization and unique functional characteristics.[12] However, individual brain regions also frequently interact with numerous other regions throughout the brain.[2] These macroscopic circuits are recurrently connected via direct projections, multi-synapse loops, and more widespread, indirect effects such as neuromodulator release.[13] Consequently, much of the brain is active during even simple behaviors that could, in theory, be mediated by only a smaller subset of regions.[14–16] Deriving an understanding of the neural basis of behavior requires consideration of the distributed nature of brain-wide activity. However, despite the prevalence of large-scale, multi-region datasets afforded by modern experimental techniques, we lack a comprehensive, unifying approach to infer brain-wide interactions and information flow. Here, we introduce Current-Based Decomposition (CURBD), a computational framework that leverages recurrent neural network (RNN) models of multi-region neural recordings to infer the magnitude and directionality of the interactions between regions across the brain. While most neural data analysis and dimensionality reduction techniques[17] describe the output of neurons (e.g. spiking activity), CURBD reconceptualizes the activity of a neural population in terms of the inputs driving the neurons[18,19]. We first introduce the conceptual advantages of CURBD and validate the method using simulated datasets where ground truth multi-region interactions are known. We then apply CURBD to several experimentally obtained datasets including calcium imaging data collected from three regions in head-fixed larval zebrafish[7,20] and four regions of mice while running,[8] and electrophysiological recordings from three cortical and subcortical regions in the rhesus macaque during a Pavlovian conditioning task[9] and four brain regions of human participants during a memory retrieval paradigm.[10] These examples highlight the widespread applicability of CURBD for inferring multi-region interactions from a range of neural datasets across different species.

Current-based decomposition of multi-region datasets using recurrent neural networks. CURBD operates on the fundamental premise that the exchange of currents between active units in a recurrently-connected neural network can be precisely estimated by knowing the activations and the weights connecting pairs of active units in the network. In a single-layer feedforward network, the currents driving a single target unit can therefore be viewed as a weighted sum of the activity of the "source" units (**Fig. 1a**). Mathematically, these weights correspond to interaction strengths, summarized by a vector with each source unit represented as a single entry in the vector. However, neural circuits in biological brains are typically intricately and recurrently connected[2] through both feedforward and feedback interactions. This feature prompted common use of RNNs to model their computational functions,[21,22] although RNNs have other advantages, such as their ability to generate time-varying activity patterns[19,23] and their trainability.[22,24] RNNs trained to produce desired behaviors[25–27] and tasks[6,28–32] or match neural data[2,6,7] (or both[18,33]) can also be reverse-engineered to generate hypotheses for how biological neural circuits could implement similar functions.[34,35] As in the single-layer example above, the activity of any unit in an RNN can be computed as a weighted sum of the activity of all other units in the network, which are the sources of its input (**Fig. 1b**). The interactions within all pairs of units in the network, which give rise to the dynamics produced by the network, can thus be described compactly using a single "directed interaction" matrix quantifying the magnitude and type (excitatory or inhibitory) of the interactions.

Given the high degree of recurrent connectivity within and between regions, interactions between active neurons in different brain regions can be implemented through an RNN framework.[2] To implement CURBD, we model brain-wide circuitry as multiple inter-connected RNNs forming a "network of networks".[2] The activity of units in each "region" of this RNN is shaped by excitatory and inhibitory "source currents" from all regions that provide inputs, including those from recurrently connected units within the same region. If the connectivity relating these networks is known[18,36]—or in our case, inferred from the network model fit to time-series data—then the source currents into a target region from any other region can be estimated using only the corresponding submatrix of the directed interaction matrix and the activity of the source region (**Fig. 1c**). When summed, these constituent source currents reconstruct the total activity observed in each neuron in the region. However, CURBD allows the total activity of a specific region to be decomposed into a set of source currents from all other regions that interact with it. Estimating population-wide source currents at this scale—from multiple regions up to potentially brain-wide interactions[7]—moves beyond simple correlations in activity between pairs of interacting regions,[37–40] giving an unprecedented view into multi-region interactions.

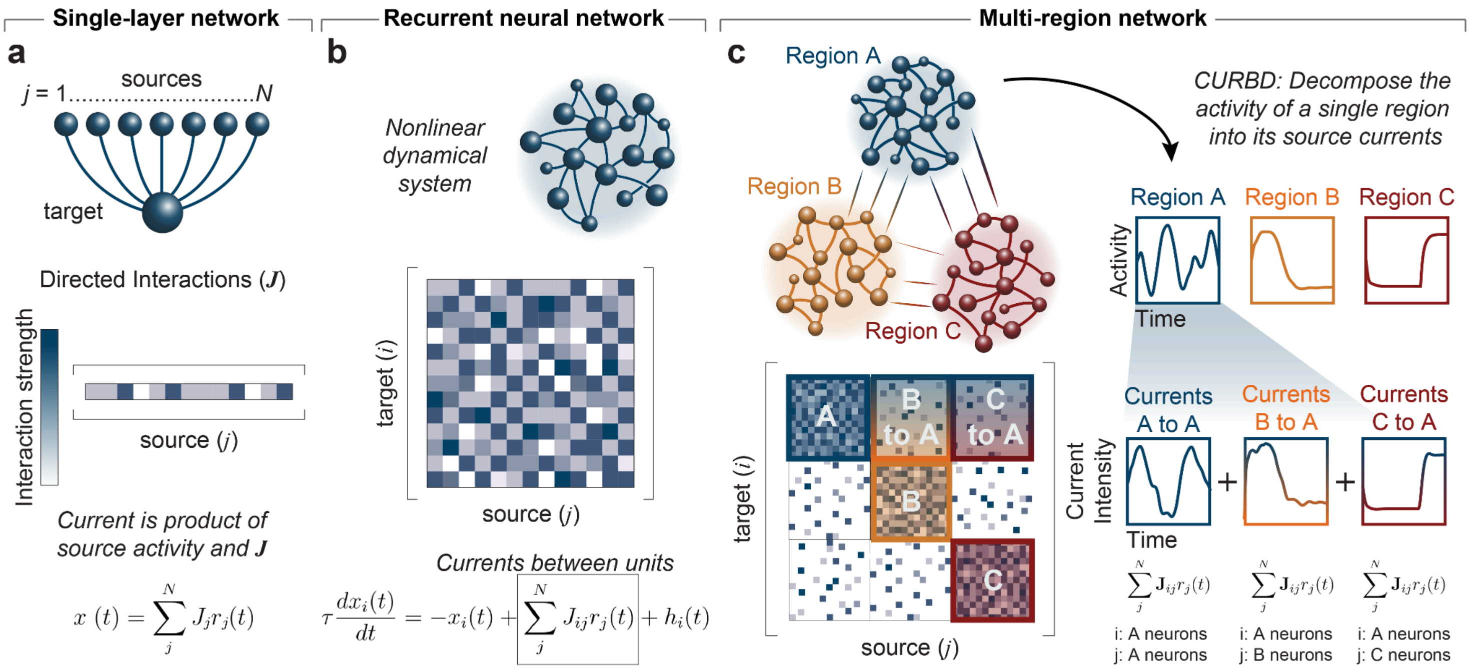
Single-layer network
Recurrent neural network
Multi-region network
a
b
c
sources
j = 1............................N
target
Directed Interactions ($\boldsymbol{J}$)
Interaction strength
source (j)
Current is product of source activity and $\boldsymbol{J}$
$x(t) = \sum_{j}^{N} J_j r_j(t)$
Nonlinear dynamical system
target (i)
source (j)
Currents between units
$\tau \frac{dx_i(t)}{dt} = -x_i(t) + \boxed{\sum_{j}^{N} J_{ij} r_j(t)} + h_i(t)$
Region A
Region B
Region C
CURBD: Decompose the activity of a single region into its source currents
Region A
Region B
Region C
Activity
Time
target (i)
source (j)
A
B to A
C to A
B
C
Currents A to A
Currents B to A
Currents C to A
Current Intensity
Time
+
+
$\sum_{j}^{N} \mathbf{J}_{ij} r_j(t)$
$\sum_{j}^{N} \mathbf{J}_{ij} r_j(t)$
$\sum_{j}^{N} \mathbf{J}_{ij} r_j(t)$
i: A neurons
j: A neurons
i: A neurons
j: B neurons
i: A neurons
j: C neurons

Fig. 1 | Current-based Decomposition (CURBD) of multi-region interactions using recurrent neural networks. a, In a single-layer network, each source unit connects to a target unit with a directed interaction weight given by the vector *J*. The activity of the target unit can be derived based on its source current, a weighted combination of the source activity (*r*) multiplied by the corresponding excitatory or inhibitory directed interaction weight. **b,** In recurrent neural network (RNN) models, each unit is driven by inputs from the other units, but also sends outputs to those same units. Thus, the directed interactions are summarized by a matrix J with each column containing the weights of a source unit and each row those of a target unit. **c,** Neural circuits can be modeled as a 'network of networks', i.e., with interconnected but distinct regions. The directed interactions governing multiple regions are still summarized by a single matrix J where submatrices along the diagonal correspond to within-region interactions, and off-diagonal submatrices correspond to interactions between different regions. As in the single-layer network in Panel a and single-module RNN in Panel b, the currents driving each target unit can be viewed as the weighted sum of source activity from each of these submatrices. By multiplying the weights in each submatrix by the source activity of each region individually, we can decompose the total activity of any region (e.g., Region A) into the constituent source currents of the total activity.

Implementation of CURBD. CURBD is based on the directed interaction matrix, **J**, which we use to infer source currents. This matrix estimates the effective strength and type—excitatory (J_{ij}>0) or inhibitory (J_{ij}<0)—of interactions between all pairs of active neurons, both within and across regions, that give rise to the experimentally-observed neural dynamics. This matrix dictates the entire neural dynamical system over time, and captures the stability as well as the population-wide covariance of the activity (**Fig. 2c**). Since a matrix capturing both the stability and structure of multi-region interactions is intractable through experimental measurements alone, we employ data-constrained Model RNNs to infer the directed interaction matrix from multi-region experimental data obtained from behaving animals (**Fig. 2a**). We first initialize a Model RNN with random connectivity (see Methods). These Model RNNs typically contain a number of units equal to the number of neurons available in the dataset to be modeled, but both larger and subsampled variants can be employed.[6] Each model unit is then assigned to one neuron in the experimental dataset during training. The goal is to learn a directed interaction matrix such that, after training, the Model RNN autonomously reproduces the time-series activity of the recorded neurons given only their initial state. At each time step, the output activity at the current time point *t* in the Model RNN, *r*[*t*], can be computed based on the sum of the previous state of population activity—*r*[*t-1*], a nonlinear function ϕ of *x*[*t-1*], the currents driving the RNN units—weighted by **J** (see Equations 4, 5 in Methods). By the above notation, we consider the activity of units as a function of time, and we use square brackets to denote that the time domain is discrete in experimental data and model implementations. For theoretical derivation where we regard the activity to be a continuous variable, we will use parentheses (e.g., *r*(*t*)) to denote the continuous function.

Training proceeds iteratively[6,7,25] (**Fig. 2b**; Methods) to minimize the instantaneous error between the activity of each experimentally recorded neuron ($a_i[t]$) and the activity of its corresponding Model RNN unit. The error can be computed in the space of currents driving the RNN units ($\mathbf{x}[t]$*;* current-based learning) or the output rates resulting from these currents ($\mathbf{r}[t]$*;* rate-based learning), where $\mathbf{r}[t]$ denotes a vector whose *i*th element is $r_i[t]$. At each time step of the learning process, the directed interaction matrix **J** is updated by $\Delta\mathbf{J}$, a function of this error (Equation 1; see Methods). Thus,

$$\mathbf{J}[t] = \mathbf{J}[t-1] - \Delta\mathbf{J}[t]_{,} \tag{1}$$

where the update is given by,

$$\Delta\mathbf{J}[t] = f(\mathbf{r}[t] - \mathbf{a}[t]) \text{ or } f(\mathbf{x}[t] - \mathbf{a}[t]) \tag{2}$$

Note that the Model RNN can be trained either based on trial-averaged neural data aligned on relevant events[6] (as in the monkey and human datasets in Results section) or, when large numbers of simultaneously-recorded neurons are available, using single-trial or even continuous time-series data (as in the mouse dataset in the Results section). At this stage, to train the Model RNN, we do not need to make any assumptions about the identity of the modeled neurons—such as which brain region they belong to, cell type, or cortical layer. The Model RNN instead fully learns a single dynamical system which reproduces the entire time series of multi-region neural data obtained experimentally using just an initial condition. In essence, after training we obtain an *in silico* model of the recorded brain regions that recapitulates the experimentally recorded multi-region data, but with crucial advantages (**Fig. 2c**): i) after training, the Model RNN generates realistic patterns of neural activity consistent with experimental data;[6,7,18] ii) training quenches the spontaneous chaotic dynamics of the randomly initialized network,[25] ensuring that the trained networks are reproducible; and iii) the trained network model contains the directed interaction matrix that CURBD leverages to infer the currents between recurrently connected units, both within- and across multiple interacting brain regions. Since the Model RNN is trained to directly reproduce time-series neural data, this directed interaction matrix is an estimate of the effective functional interactions between each recorded neuron. These functional interactions can be distinct from true anatomical connectivity since they can include long-range or polysynaptic pathways, and indirect effects such as neuromodulator release. Consequently, the currents in the RNN can be thought of as a functional estimate of the information exchanged between neurons in the recorded dataset; whether or not currents inferred by the CURBD method represents a measurable indicator of physiological currents such as postsynaptic potentials remains to be tested.

The current $I_i(t)$ received by any one target unit $i=1$, 2, …, N can therefore be viewed as the sum of the activity of the N source units scaled by the respective interaction weights between the source units and the target unit.[18,41]

$$I_i(t) = \sum_{j}^{N} J_{ij} r_j(t) = J_{i1} r_1(t) + J_{i2} r_2(t) + ... + J_{iN} r_N(t) \tag{3}$$

All of the constituent source currents in the previous equation sum to reconstruct the full activity of units in the target region. In this paper, we focus on currents exchanged between brain regions. By restricting the summation in Equation 3 to source units from a specific region, we can isolate the currents into the target region from a specific source region (**Fig. 1c**). In practice, based on labels applied to each experimentally recorded neuron, the matrix **J** can be broken into M^2 submatrices, where M is the number of regions (or other relevant clusters) identified in the experimental dataset, corresponding to all pairs of source/target region interactions in the region (**Fig. 1c**). Note that in this paper we assume that the region identities for each neuron are known *a priori* through anatomical labeling or other forms of clustering. This separation of currents can thus be considered a decomposition of the activity of the target-region neurons based on the relative source current contributions of each source region. These source currents inferred by CURBD can be powerful tools to analyze existing neural data and help guide new experiments to dissect interactions. The method provides directional estimates of the excitatory and inhibitory interactions between large numbers of regions. Direct analysis of the characteristics (e.g., strength, type, or timing) of the disparate current inputs can help identify new functional subcircuits in multi-region data, and unveil inter-region communication mechanisms that may be inaccessible through experimental measurements alone.

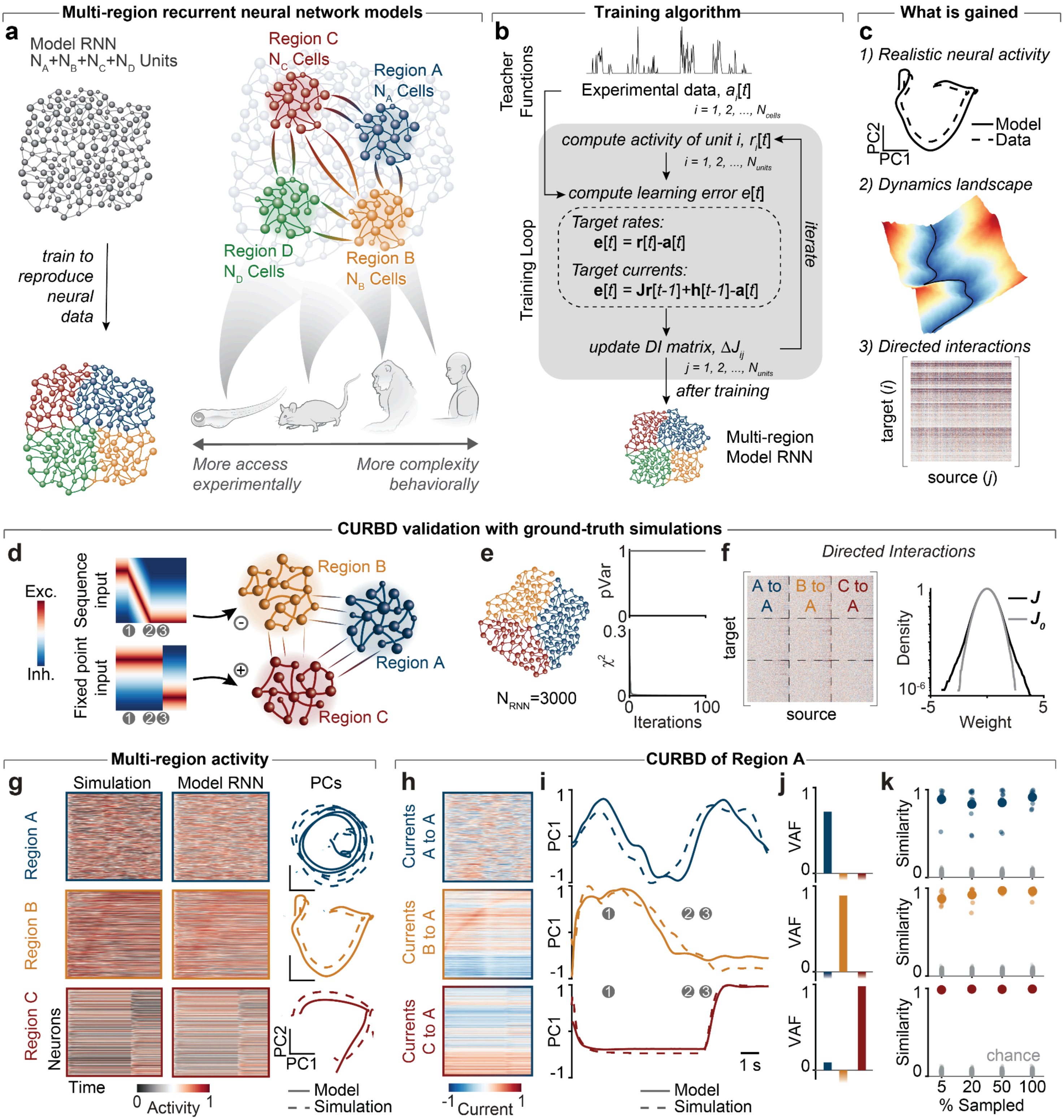

Fig. 2 | Data-constrained multi-region RNN design, training procedure, outcomes, and validation. a, CURBD is a flexible framework to understand datasets of varying complexity and experimental accessibility. CURBD is implemented through a Model RNN constrained from the outset by experimental data. Neural data from experiments in behaving animals (here, zebrafish, mice, monkeys, and human participants) are segmented into modules such as brain regions. A Model RNN is constructed such that each unit is trained to match a single experimentally recorded neuron from the full dataset of neural population activity from multiple interacting regions. **b,** Training occurs iteratively, where the connectivity matrix **J** of the Model RNN is modified over time until the activity of the RNN units match the experimental data. Training can be performed in the space of the rates of the RNN units or the currents driving the units. **c,** This approach has several advantageous outcomes. 1) The model, after training, exhibits realistic neural dynamics consistent with experimental data. 2) The dynamical landscape, which goes beyond capturing covariance structure of activity, can be further analyzed after training. 3) The directed interaction matrix inferred by the trained Model RNN gives unique insight into the functional connectivity responsible for the observed dynamics in the data, including strength and type (e.g., excitatory or inhibitory weights and unidirectional or feedback projections between regions) of interactions, both within and across regions. **d,** We generated three RNNs representing three distinct brain regions. Region B was externally driven by a sequentially active population, Region C was externally driven by a population generating fixed points, and Region A was driven only through interactions with Regions B and C. Time points 1-3 denoted by gray circles represent key time points in these external inputs. **e,** We fit the Model RNN to the time-series data of all three regions comprising the generator model, reaching high variance explained (pVar=0.99) and low training error (χ^2=7.4x10^{-6}). **f,** (Left) Training resulted in the directed interaction matrix describing the interactions within and between regions. (Right) Normalized distribution (log scale) of all weights in the Model RNN directed interaction matrix (**J**, black) compared with the randomly initialized matrix ($\mathbf{J_0}$, gray) **g,** After training, the Model RNN accurately reproduced the single-unit activity of all three simulated regions (left), as well as the population trajectories in the leading principal components (PCs; right; solid lines: model; dashed lines: simulation). **h,** The individual current sources into Region A showed qualitatively similar activation patterns to those in the source regions, even though these patterns were not apparent in the population activity of Region A. **i,** The dominant PC of each source current (solid lines) and the ground truth inferred through the known connection weights in the generator model (dashed lines). Gray circles are the same key time points as in Panel a. **j,** Variance Accounted For (VAF) for the first PC of each source current compared to the three ground truth source currents (VAF_{AtoA}=0.72; VAF_{BtoA}=0.89; VAF_{CtoA}=0.98). **k,** We repeated the simulation while randomly sampling different subpopulations of the 1000 unit networks ranging from 5 to 100% of the total units. Note that the 100% mark represents repeated training runs of the same neurons with different random initializations of $\mathbf{J_0}$. We quantified CURBD performance using a “similarity metric” of the current dynamics. Small dots represent the similarity obtained by 10 different repetitions; large dots represent the average. Gray dots show the lower-bound obtained from 100 random shuffles of the ground truth dynamics.

Validation of CURBD on ground truth datasets. Since CURBD was designed to infer unobservable interactions in experimental datasets, we first validated the method in synthetic datasets where the ground truth inter-region currents are simulated. We created a generator model comprising three RNNs, initialized in the chaotic spontaneous activity regime,[42] representing three distinct ‘regions’ and connected through sparse inter-region connection weights (**Fig. 2d**). Region B was externally driven by a sequential external input, Region C was externally driven by fixed-point inputs, and Region A was driven only through sparse, inter-region recurrence with Regions B and C. We selected the parameters of this model such that the simulated activity in Region A was visibly chaotic without any clear representation of either the sequential or the fixed-point external input driving the regions connected to it (**Fig. 2g**).

We trained a single Model RNN (**Fig. 2e**) to match the simulated data from the generator model described above (**Fig. 2f,g**). Since the ground truth data was generated by known currents in the three-region Generator RNN, we opted to train the Model RNNs using the current-based learning rule described above. We hypothesized that CURBD would accurately infer the inputs that Region A receives indirectly from source Regions B and C, despite the chaotic nature of the population activity observed in Region A. Using the respective submatrices of the directed interaction matrix inferred from the trained Model RNN (**Fig. 2f**), we decomposed the activity of Region A into the currents from each source region. These currents reflected qualitatively similar activation patterns to those in the source regions (**Fig. 2h**)—sequential and fixed point, respectively—even though these patterns were not apparent in the population activity of Region A. Since the true connectivity of the simulated network was known, we also computed the ground truth currents into Region A. We summarized the population-wide currents from each source using principal components analysis (PCA) and compared the CURBD output to ground truth using Variance Accounted For (VAF). CURBD accurately reconstructed each current source driving Region A (**Fig. 2i**). We then compared the performance of CURBD to canonical correlation analysis (CCA), which has been used to identify individual subspaces that capture the shared dynamics.[43,44] We found that CCA did not accurately capture the ground truth currents (**Fig. S1c,d**) due to the recurrence in the Generator network.[45]

We next adapted this idealized ground-truth simulation to test the practical limits of CURBD and identify whether any key limitations exist. In real datasets, despite explosive advances in neurotechnologies,[15,16,46–49] experimenters typically only have access to a small percentage of neurons in a given region,[50] with the sampling ratios falling precipitously with the size of the brain. We repeated the simulation to test whether CURBD is effective when the brain regions are partially sampled by training the Model RNN based on targets from only a small fraction of units. We computed a "similarity metric" that could quantify the similarity of the dynamics of the currents inferred by CURBD dynamics across the range of different sampling ratios.[27,43,51] CURBD accurately estimated the current dynamics even when the network was highly undersampled, as low as 5% of the population (**Fig. 2k**). We then designed a second simulation to explore more challenging scenarios for CURBD. We simulated two reciprocally connected RNNs, each receiving sinusoidal inputs of different frequencies (**Fig. S1e-k**). Since the sinusoidal inputs can mix with the ongoing chaotic dynamics in recurrent networks,[42,52] they provide a more challenging paradigm to assess CURBD. We found that CURBD was most effective when the intrinsic dynamics of the two RNNs were distinct—resulting from differences in external drive, internal structure or connectivity, or both—with sparse inter-region connectivity (**Fig. S1k**). These results from validating CURBD on simulated ground truth datasets illustrate that CURBD can infer unobserved source currents between multiple brain regions under a range of experimentally realistic conditions.

Comparing different learning rules in models. The ground truth datasets above were based on currents flowing through simulated RNNs. Accordingly, we trained the Model RNN to target either the current inputs to each cell (current-based learning).[6,7] However, many real datasets consist of recordings or imaging studies reflecting the firing rate or spiking outputs of neurons across different brain regions. Furthermore, current-based fitting makes assumptions about the nonlinearity in the biological neuron—usually a specific type of mathematical function[53]—which needs to be inverted to extract from the experimentally recorded output a target function with units consistent with a currents-based representation of error (see Materials and Methods of Ref. [6]). Rate-based fitting makes fewer assumptions and could therefore be used to infer a more unbiased estimate of the interactions based on recorded or modeled activity. Therefore, we ask whether real neural datasets are better fit by targeting the resulting output rates of each experimentally recorded or imaged neuron (rate-based learning) (see STAR Methods). We performed this comparison on an experimental dataset with a large number of simultaneously-recorded neurons (**Fig. S2 and S3**). In brief, we used optical recording of fluorescence from genetically encoded calcium sensors to simultaneously track the activity of thousands of neurons in five behaving larval zebrafish expressing GCaMP6.[7,54] We fit Model RNNs to reproduce the neural recordings from both the telencephalon and thalamus of each fish (**Fig. S2C**). We compared the Model RNN fits obtained using both rate-based and current-based learning (see Methods) for five training runs starting from different randomly initialized states. In this experiment, all initialization and input parameters were identical for both learning rules. We found that the two learning rules gave comparable performance. However, for our purposes, we identified two advantages to rate-based learning: 1) higher explained variance by the Model RNN, indicating better fit for neural data (**Fig. S2f**); and 2) greater consistency across multiple, separately initialized runs (**Fig. S2g**). We therefore employ rate-based learning for the following applications of CURBD to neural datasets collected experimentally from zebrafish, mice, macaques, and humans.

Decomposition of three regions from larval zebrafish during a behavioral challenge paradigm. Here, we use CURBD to infer mechanisms underlying responses to a behavioral challenge paradigm in larval zebrafish. In brief, larval zebrafish (6 days post fertilization) expressing GCaMP6s[7,54] were partially embedded in agar and presented with inescapable electric shocks (**Fig. 3a**). During this paradigm, we imaged calcium activity from three regions: the habenula, the raphe nucleus, and the telencephalon, a large, multi-faceted, cortex-like region whose activity relates to many aspects of behavior[55] (**Fig. 3b**). Previous work demonstrated that this paradigm induced depression-like behaviors in the fish.[7] Initially, the fish adopted an "active coping" strategy with escape-like tail movements (**Fig. 3a**). However, since the shocks were inescapable, as stress accumulated the fish gradually progressed to a “passive coping” strategy with minimal movement that minimized energy expenditure, akin to learned helplessness in depression.[56,57] This behavioral state transition is mediated by interactions between the habenula and the raphe nucleus,[7] though it remains unclear how habenula is recruited during this process.

We applied CURBD to infer the multi-regional inputs that drive habenula to cause the behavioral state transition. We fit Model RNNs to single-neuron calcium traces taken from the three imaged regions. We then performed CURBD to isolate the current sources driving habenular activity (**Fig. 3c**). The Model RNNs converged accurately and consistently for all five fish in the shocked cohort (**Fig. 3d-g**). We then studied the relationship between the fish’s behavioral response and the temporal dynamics of each current source to the Habenula during the behavioral challenge paradigm (**Fig. 3h**). We compared the behavior of the shocked cohort to a control group which received no shocks and found that the shocked fish gradually decreased their tail movements as they adopted a passive coping strategy (**Fig. 3i**). We found a striking difference between inputs from different regions, with initial habenular activity driven primarily by inputs from the raphe nucleus during the active coping phase (**Fig. 3j**). Later, as the fish transitioned into passive coping, we observed ramp-like recruitment of habenular neurons by the telencephalon and intra-habenular currents. This separation between inputs from the raphe nucleus and habenula was not present in the cohort of control fish. These current profiles indicate two different timescales of habenular inputs corresponding to the distinct behavioral states.

The Model RNNs give us access not only to the estimated current dynamics between regions, but also the strengths of interactions via the weights of the interaction matrix. We analyzed the distribution of inferred interaction weights between the raphe nucleus and habenula in the early phase of shocks (active coping) and the late phase (passive coping) (**Fig. 3k**). We found that the early phase, which was marked by large inputs from raphe nucleus to habenula, had no change in weights compared to the pre-shock baseline. Only during the late phase, after the behavioral state transition occurred, did we see a corresponding change in interaction weights. The joint insights from the current dynamics and interaction weights show that the habenular responses to the behavioral challenge are driven by two distinct mechanisms operating on two different timescales.

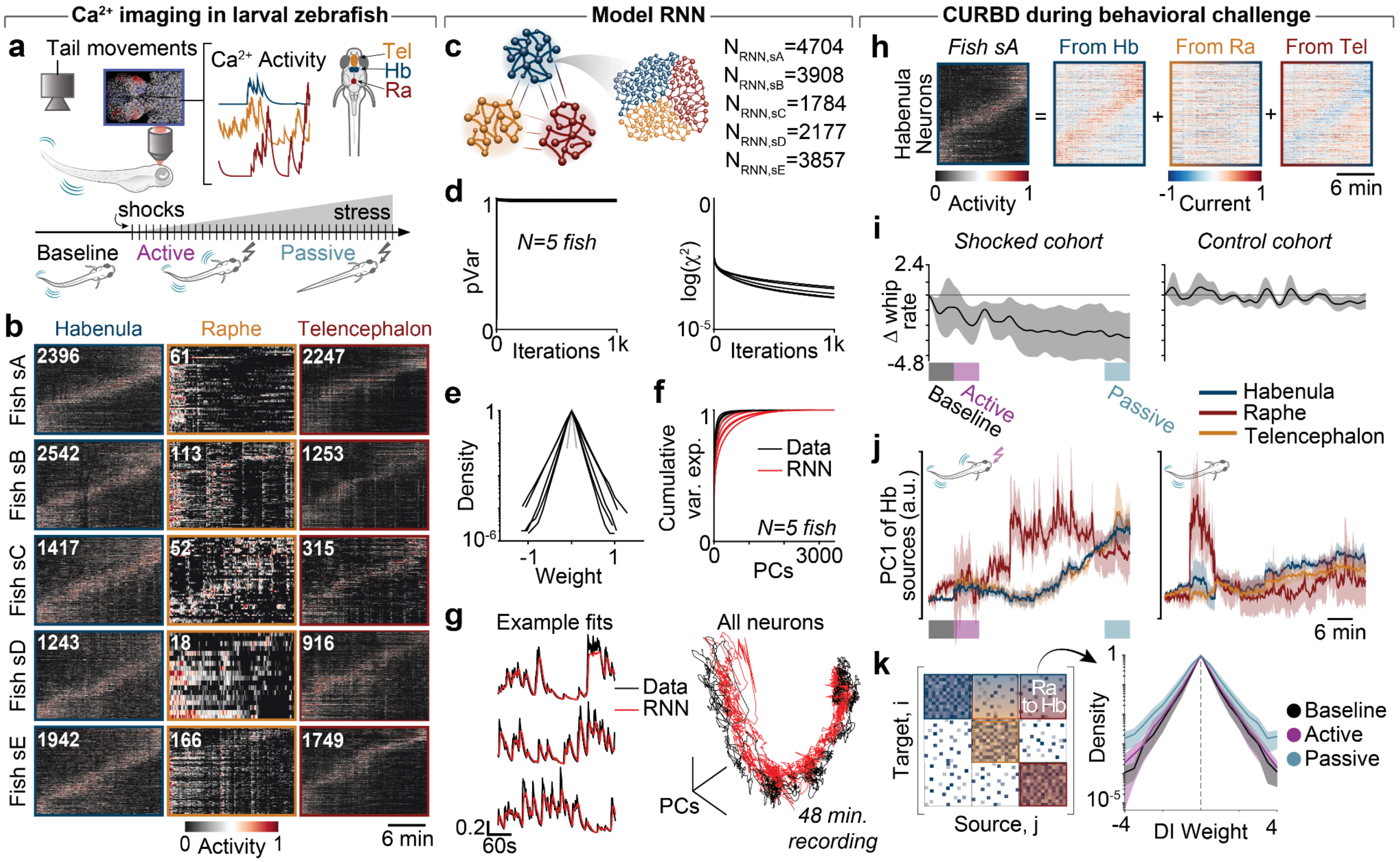

Fig. 3 | Isolating source currents into the habenula from multi-region calcium recordings in larval zebrafish. a, (Top) We recorded multi-regional neural activity with Ca^{2+} imaging in larval zebrafish expressing GCaMP6s and simultaneously recorded tail movements using a high-speed camera. (Bottom) The fish experienced gentle but inescapable electric shocks inducing stress responses which, over time, shifted from active coping (escape movements) to passive coping (minimal energy expenditure). **b,** Neural population activity from the three regions during 48 minutes of recording in five shocked fish. Text inset denotes the number of recorded neurons in each region. **c,** We fit Model RNNs sequentially in 6 minute epochs to all neurons from the three regions. **d,** Proportion of variance explained (pVar, left) and training error (χ^2, right) for the RNNs trained to match data in the Baseline epoch of all five shocked fish. **e,** Normalized distribution of interaction weights (log scale) for the five Model RNNs before (gray) and after (black) training. **f,** Cumulative variance explained by all PCs for the neural data (black) and Model RNNs (red). Each line represents one of the 5 fish. **g,** (Left) Comparison of example single neuron fits from an example fish (Fish sA) for the neural data (black) and Model RNN unit (red). (Right) The leading three Principal Components capturing the covariance across all neurons in the three regions for the neural data and Model RNN over 48 minutes of recordings in the example fish. **h,** CURBD decomposition for Fish sA. (Left) Heatmap of RNN unit activity for the Habenula. (Right) Heatmaps of current decomposition for each of the three source currents into Habenula. **i,** Change relative to the baseline epoch of tail whip movements over time in the shocked cohort (left) and a control cohort who underwent the same experiment but did not receive shocks (right). Data presented as mean ± s.e.m. across the 5 fish in each cohort. Blocks at the bottom denote the baseline (black), active coping (magenta), and passive coping (cyan) epochs. **j,** Temporal dynamics of the first Principal Component capturing inputs to habenula from habenula (blue), raphe nucleus (red) and telecephalon (yellow). Data presented as mean ± s.e.m. across the 5 fish in the shocked (left) and control (right) cohorts. **k,** Normalized distribution of weights in the sub-matrix capturing raphe to habenula interactions inferred by the Model RNN for the baseline (black), active (magenta), and passive (cyan) epochs. Data in each epoch presented as mean ± s.e.m. across the 5 fish in the shocked cohort.

CURBD untangles brain-wide currents during spontaneous movement in mice. Here, we demonstrate that CURBD untangles behaviorally relevant source currents from a large-scale, multi-region calcium imaging dataset in mice. Mice expressing GCaMP6s[5] were allowed to run freely on top of an air-supported ball in complete darkness (**Fig. 4a**). Using a large field-of-view two-photon microscope,[58] we imaged neural activity simultaneously from four regions (**Fig. 4a,b**): primary visual cortex (V1), secondary motor cortex (M2), posterior parietal cortex (PPC), and retrosplenial cortex (RSC). Together, these regions are thought to contribute to a brain-wide circuit governing complex functions such as navigation, decision-making, and movement.[59–61] Mice exhibited spontaneous bouts of running behavior, measured through rotations of the air-supported ball (**Fig. 4c**), with complex patterns of neural activity observed across all four brain regions during these bouts. Consistent with recent studies (e.g., Ref. [15,16,62]), we observed a high degree of activity even in V1 despite the fact that the mice received no visual input, highlighting the distributed nature of behavior-related activity throughout the brain.

We hypothesized that CURBD could isolate different sources of behaviorally-relevant information in regions such as V1, in which recorded data is often high dimensional, multiplexed across multiple spatial and temporal scales, and results from interactions spanning multiple brain regions.[14,16,62] To test this, we first trained Model RNNs to reproduce the neural data from the four recorded regions (**Fig. 4d-f**). Applying CURBD, we identified strikingly different patterns of excitation and inhibition during running bouts for the sixteen source currents (**Fig. 4g and S5a,c**). Our analysis focused on the currents into V1 since we wanted to isolate sources of running-related signals. We computed the relative variance explained by each source current of the full V1 population activity (**Fig. 4h**). We originally predicted that currents from M2 and PPC, which are closely involved in planning and producing behavior,[59,63] would increase during bouts of running. However, we saw no clear relationship between the variance captured by each source current and the running speed. Instead, source currents into V1 from each of the four brain regions occurred in similar proportions, indicating that balanced currents flow steadily between the four regions even while the mice are at rest.

To analyze the population-wide dynamics of these currents, we computed separate low-dimensional neural manifolds[17,64,65] spanning each source current using PCA. The trajectories within these manifolds (**Fig. 4i and S5b,d**) capture the dominant dynamics of each source current into the target region. Studying the dynamics of the source currents into V1 (**Fig. 4i**), we observed that while M2 and PPC currents showed large deviations in their trajectories during running bouts, RSC to V1 currents did not change much. These observations suggest that information related to locomotory behavior may arrive in V1 selectively from M2 or PPC. To test this quantitatively, we built linear decoders predicting running speed based on each source current to V1. Comparing decoders trained using the currents into V1, we found that the currents from M2 contained the most information about running speed (**Fig. 4j-l and S6**). These results illustrate the potential of CURBD to untangle the complex, multi-regional interactions underlying behavior using RNNs based on multi-region calcium imaging data.

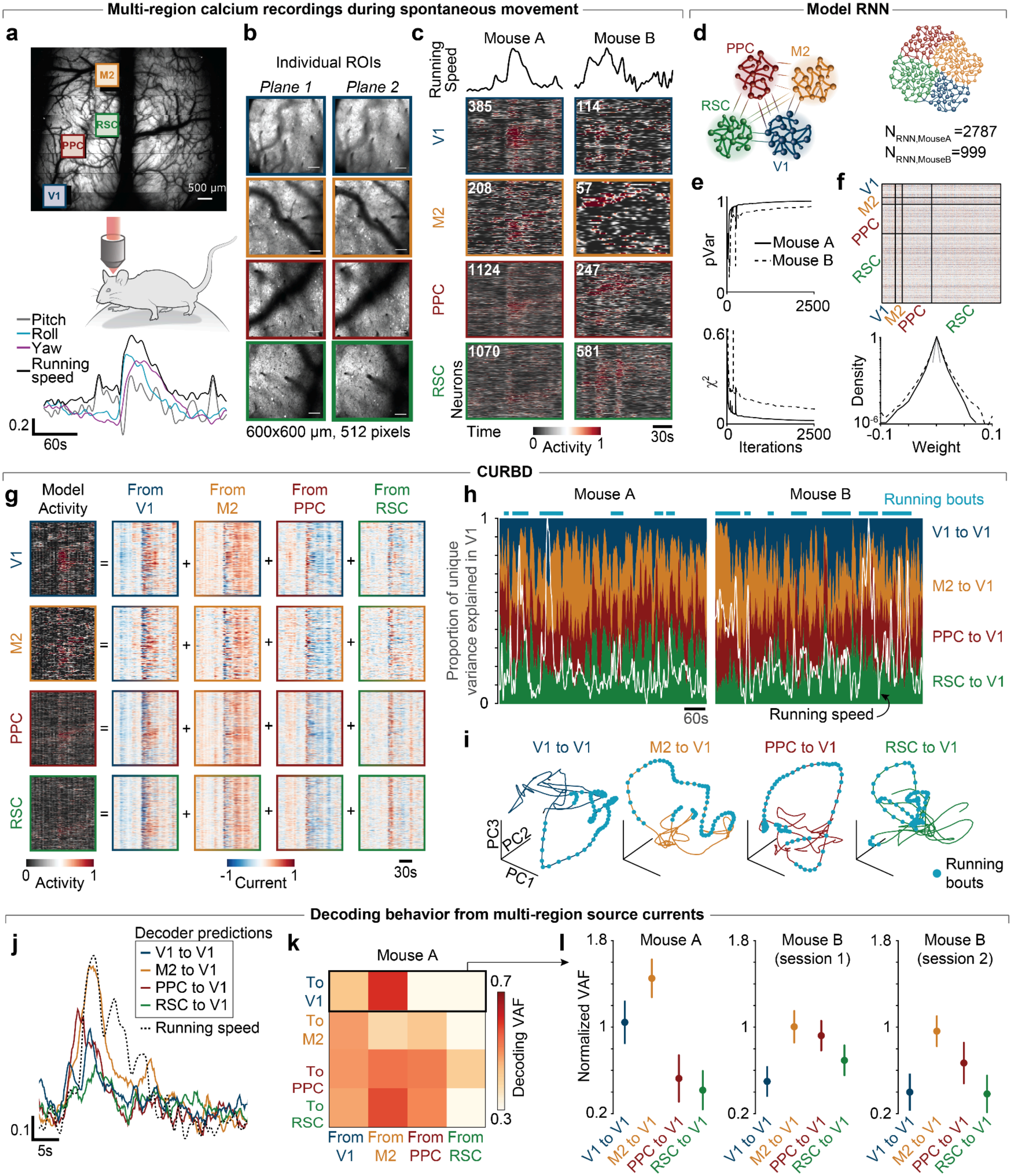

Fig. 4 | Isolating source currents from multi-region calcium recordings in mice. a, We recorded neural activity from four brain regions in two mice expressing GCaMP6s. Mice were head-fixed on an air-supported ball in complete darkness. Running was tracked using sensors recording the pitch (gray), roll (cyan), and yaw (magenta) velocities of the ball; the total magnitude of the three signals, combined, is summarized as running speed (black). **b,** We imaged two planes from each brain region (regions of interest, ROIs). **c,** Behavior and neural population activity from the four regions during a brief period of spontaneous running for two mice. Text inset denotes the number of recorded neurons in each region. **d,** We used ten consecutive minutes of recordings to fit a Model RNN for both mice. **e,** Proportion of variance explained (pVar) and training error (χ^2) for the RNNs trained to match data from Mouse A (solid lines) and two sessions from Mouse B (dashed lines). **f,** (Top) Example directed interaction matrix for Mouse A. (Bottom) Normalized distribution of interaction weights (log scale) for the three Model RNNs before (gray) and after (black) training. **g,** CURBD decomposition for Mouse A. (Left) Heatmaps of RNN unit activity for the four regions. (Right) Heatmap of current decomposition for each of the sixteen source currents capturing all possible inter-region interactions. **h,** The proportion of unique variance explained by each source current of the total V1 activity. Running speed is overlaid in white, and cyan lines indicate running bouts. **i,** V1 source current trajectories in the three leading PCs for Mouse A. Cyan dots denote time points at which the mouse running speed was above a threshold ball speed. **j,** We used linear decoders to predict running speed from each source current. Example decoding predictions (colored lines) of the measured running speed (dashed line) for the four source currents into V1 for Mouse A. **k,** VAF for all sixteen source current decoders for Mouse A. **l,** Decoder performance (mean and standard deviation across 1000 random cross-validated test sets; see Methods) for source currents into V1 for Mouse A and two sessions from Mouse B.

CURBD applied to spiking data collected during Pavlovian conditioning in monkeys. We next applied CURBD to population spiking activity acquired by electrophysiological recordings in macaques. Many multi-region population datasets obtained by electrophysiology in primates are constructed using "pseudopopulations" where neurons recorded at different times are pooled together by averaging across repetitions of the same condition. We thus aimed to demonstrate the applicability of CURBD in inferring currents from such pseudopopulations. We obtained neural data from two monkeys (*Macaca mulatta*) performing a Pavlovian conditioning task (**Fig. 5a**). The monkeys learned to associate three conditioning stimuli with three different reward levels of increasing desirability: no reward, water, and juice. After a brief anticipatory delay, the monkeys received the expected reward. Using intracranial electrodes, we recorded from neurons in the amygdala, subcallosal anterior cingulate cortex (ACC), and rostromedial striatum. These regions are known to be important for reward processing and affective behaviors.[66] Since we could record only small numbers of neurons on a given session, we constructed pseudopopulations of 343 neurons for Monkey D and 199 neurons for Monkey H by averaging neural activity across all trials for each condition on each session. All three regions also displayed condition-specific, time-ordered sequence-like activity[9] (**Fig. 5b and S7**).

We trained Model RNNs to reproduce the neural data from all three regions for each monkey (**Fig. 5c and S7**). Once trained, the RNNs were able to produce neural dynamics similar to the recorded activity of the three regions, even though the neural data were not simultaneously collected. The nine source currents inferred by CURBD showed distinct dynamics for each region in the circuit (**Fig. 5d**). One notable advantage of CURBD is that it can infer directed inter-region currents to determine, for example, whether the interactions between any two regions are reciprocal or unidirectional/feedforward. Since neurons in ACC directly project to the rostromedial striatum,[67] we focused our analysis on these two regions. Intriguingly, we found that the strength of interactions between striatum and ACC were also asymmetric (**Fig. 5e and S6**). Since the inferred currents are the product of both the interaction strength and the source region activity, we further dissected the asymmetries in the interactions using the total magnitude of current between the two regions (**Fig. 5f,g**). CURBD revealed a high magnitude of currents from Striatum to ACC following the water stimulus, but no corresponding current from ACC to Striatum. However, on the juice trials, we saw strong bidirectional currents. Crucially, the currents inferred through CURBD were consistent across RNNs based on data from the two monkeys, as well as across five different random initializations of the Model RNN.

Since the pseudopopulations are constructed *post hoc*, their size and the specific neurons that are chosen for inclusion in the population can sometimes be prone to empirical biases or the specific limitations of the experiments. We tested whether CURBD infers the same population-wide current dynamics with pseudopopulations constructed by sampling different subsets of neurons from the full population. To do this, we applied CURBD to randomly subsampled the available neurons from each region to create pseudopopulations of different sizes (between 60% and 90% of the total numbers recorded). We computed a similarity metric, similar to the metric employed in the simulated and zebrafish datasets above, for each of the nine source currents, comparing the inferred currents at each sub-sampling percentage to the currents inferred when using the full population. We found a high degree of similarity in the identified currents even when using just 60% of the available neurons (**Fig. S8**). Thus, CURBD can be readily applied to pseudopopulations comprising non-simultaneous recordings, yielding robust estimates of the interactions between regions, even under different partial sampling conditions.

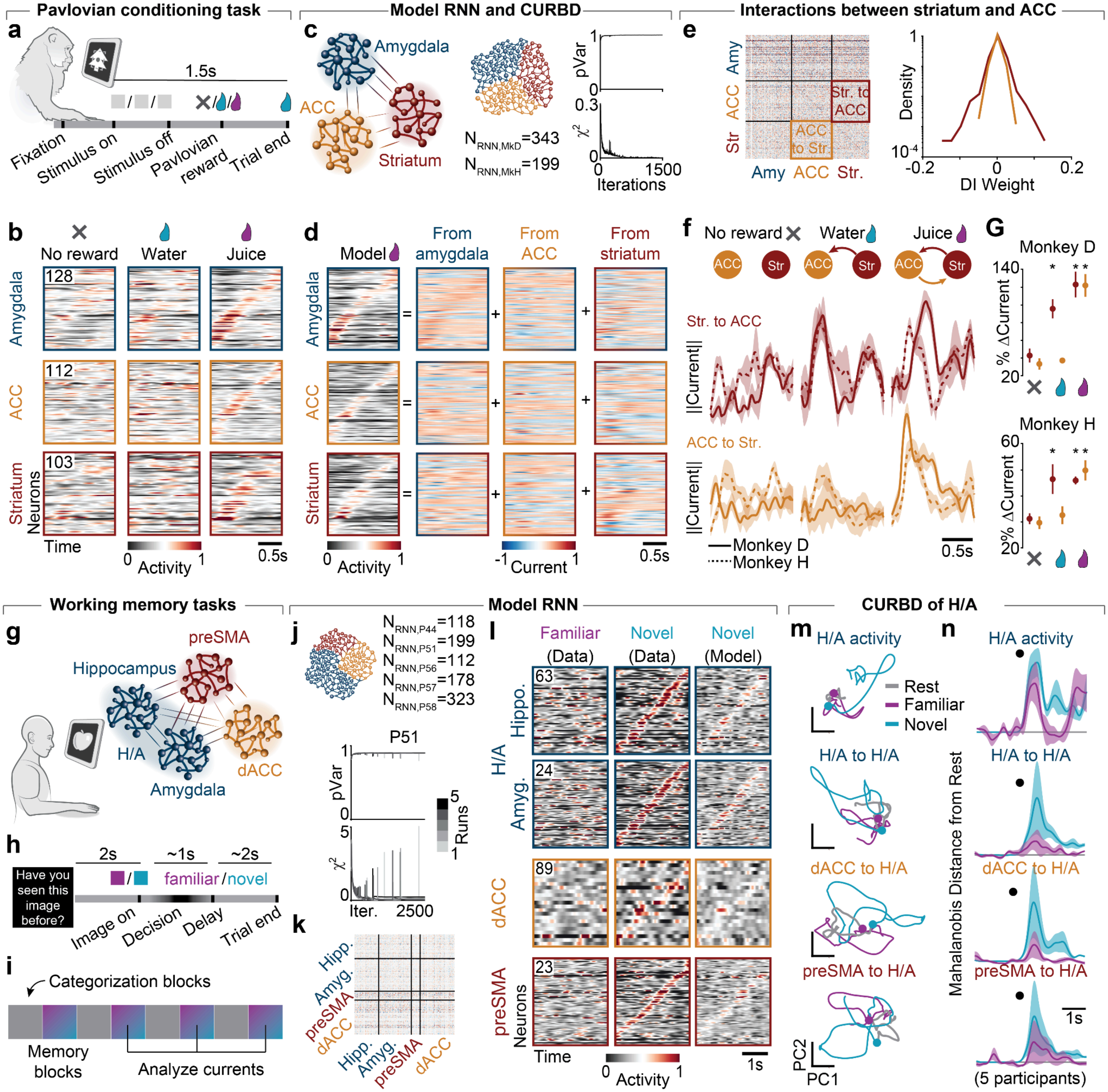

Fig. 5 | Current-based decomposition of three-region pseudopopulation recordings in primates. a, Macaque monkeys performed a Pavlovian conditioning task where one of three stimuli associated with no reward (unconditioned stimulus), water, or juice were presented for 1 second. The associated reward was delivered after a short delay (0.4-0.6 seconds) then a second water reward signified the trial end. **b,** Trial-averaged firing rates in the pseudopopulation dataset for Monkey D for the amygdala, subcallosal ACC, and striatum during the unconditioned stimulus (left, inset number denotes neuron count in each region), water stimulus (middle), and juice stimulus (right). Neurons in each region are aligned on the presentation time of the stimulus and sorted according to their time of peak activity in the juice condition. **c,** We fit Model RNNs to the three-region dataset. **d,** (Left) Schematic of Model RNN. (Right) Proportion of variance in the neural population explained by the model (top, pVar) and training error (χ^2, bottom) as a function of the number of training iterations. **e,** CURBD of activity in each region for the juice trials. Left heatmaps show the full Model RNN activity. Remaining heatmaps show the decomposition for each of the sixteen source currents capturing all possible inter-region interactions. (Left) Directed interaction matrix for an example Model RNN from Monkey D. (Right) Distribution of weights (log scale) in the striatum to ACC (red) and ACC to striatum (yellow) submatrices. **f,** Magnitude of bidirectional currents from striatum to ACC (red, top) and ACC to striatum (yellow, bottom) during presentation of the three stimuli. Solid line: Monkey D; dashed line: Monkey H. Error bars: standard deviation across five different random initializations of the Model RNNs. Schematics (top row) summarize the dominant source currents inferred by CURBD—magnitude and directionality—between the two regions. **g,** Statistical summary of the percent change in total current in the first 1s of each condition compared to the mean in the unrewarded condition. Points represent mean and lines s.e.m. *: significance at $p<0.05$, t-test. Human participants were implanted with depth electrodes to record single-unit spiking activity from neurons in the hippocampus and amygdala (combined and abbreviated H/A), pre-supplementary motor area (preSMA), and dorsal anterior cingulate cortex (dACC). **h,** Trial structure during each memory block. After a two-second baseline period, a familiar or novel image was presented. Participants reported whether they had previously seen the image (familiar) or not (novel). **i,** Each experimental session comprised eight blocks. In odd blocks the participants categorized images; in even blocks, as schematized in Panel b, the participants reported whether a presented image was novel or familiar. We used data from blocks 4, 6, and 8 to compare familiar and novel stimuli when task performance was highest. **j,** Training performance (pVar and χ^2) for Model RNNs in P51, shades of gray denote five different random initializations (runs). **k,** Directed interaction matrix of a Model RNN trained to match data from P51. **l,** Pseudopopulation activity during the memory task (Block 4) from P51 following familiar (left, inset number denotes neuron count) and novel (middle) images, and the corresponding Model RNN activity on novel trials (right) for the four regions. Neurons within each region are sorted based on the time of peak activity in the recorded data on the novel trials. **m,** Population trajectories projected onto the leading two PCs for H/A neurons during the pre-stimulus baseline period (Rest, gray) and in response to familiar (magenta) and novel (cyan) stimuli, and the source current trajectories within H/A for the two types of stimuli. Dot indicates the state at the time of stimulus onset. **n,** Mahalanobis distance from the pre-stimulus rest period computed over time for each source current into H/A. Dot indicates time of stimulus onset.

CURBD applied to single-cell spiking data from humans during memory retrieval. We next applied the method to cellular resolution, multi-region, spiking data from humans. Five participants performed a set of two memory tasks in eight interleaved blocks (**Fig. 5g-i**).[10] In the first, participants categorized images based on high level sensory features. In the second, participants were presented with an image and reported whether or not they had seen the image before. As the participants performed this task, we recorded the activity of neurons in two frontal cortical regions—pre-supplementary motor area (preSMA) and dorsal anterior cingulate cortex (dACC)—and two subcortical regions–the hippocampus and amygdala (H/A) using hybrid depth electrodes.[68] Using the same procedure as in the monkey dataset, we constructed pseudopopulations from neurons recorded from between two and five sessions in each participant. Since some datasets had data from few recorded neurons in either hippocampus or amygdala, we combined data from the two subcortical regions for later analyses.[10] Memory retrieval is believed to be mediated by interactions between frontal cortices and the H/A.[10] Our goal was to test the hypothesis that CURBD can separate currents related to the memory retrieval and memory formation within these regions. Thus, we focused our analysis on the memory task, where participants accessed their memory after viewing each image and instantiated a new memory following a novel image.

We first fit Model RNNs to the pseudopopulation datasets from each participant (**Fig. 5j-l and S9**) to estimate the directed interaction matrices. We then performed CURBD to infer the currents driving H/A following presentation of familiar or novel images. The state space trajectories in the first two PCs of the full H/A activity and each source current showed distinct current dynamics between the categorization and memory tasks after image presentation (**Fig. S10**). Within the memory task, the currents within the circuit also changed between novel versus familiar image conditions, with familiar images causing a small response in the frontal cortex to H/A currents and novel images causing a large response in all currents (**Fig. 5m**). We quantified these effects using the Mahalanobis distance from the cluster of resting state activity (**Fig. 5n**). Across all five participants, CURBD identified a substantial change in currents (relative to baseline) from preSMA to H/A when viewing familiar images, and smaller changes in the other source currents. Viewing novel images, on the other hand, caused large sustained currents throughout the whole network. These results suggest a specificity in the inter-region interactions inferred by CURBD: frontal cortex provides input to H/A during memory retrieval, while the remaining pathways are recruited following a novel image to encode new information into memory.

Discussion

Advantages of CURBD

Typical data analysis approaches study population activity from the perspective of the experimentally measured outputs from a neural circuit (e.g., action potentials through electrophysiology or calcium fluorescence signals through imaging). Using dimensionality reduction techniques,[17] we can estimate low dimensional neural manifolds[64,69] embedded in the space of total population activity. Neural manifolds are defined by patterns of coactivation between neurons in measured population activity. However, the covariance observed in neural populations is shaped by the inputs driving that population,[70] this includes extensive feedforward and feedback (recurrent) interactions within and across multiple brain regions.[4,46,47,58,71–73] CURBD offers a unique view of neural activity by decomposing experimentally measured population activity into such inferred inputs or 'source currents'. Rather than identifying a single manifold capturing the measured outputs of active units within a given region, we can use such source currents to compute separate manifolds for each source current inferred. Therefore, CURBD allows us to reconceptualize population activity as numerous manifolds embedded in the space of neural activity, each capturing the dynamics of a single, isolated source of input.

CURBD builds on and significantly extends existing approaches using models directly constrained by experimental data[2,6,7,18,34,36,74] to understand the functional architecture of neural circuits or the statistics of multi-region interactions. CURBD goes beyond this foundational work to build and analyze multi-region RNNs constrained directly by neural data from different species–here, larval zebrafish, mice, macaques, and humans. Reverse engineering such data-constrained multi-region RNNs let us consistently infer quantities that are inaccessible from measurements alone: i) the directionality, type, and magnitudes of the interactions responsible for the observed neural dynamics can be inferred through the directed interaction matrices;[7] and ii) the time-varying behavior of different inter-area currents in these diverse nervous systems using the CURBD approach presented here. CURBD addresses several gaps in commonly applied computational approaches for analyzing the large-scale experimental data enabled by explosive new technologies for monitoring large scale neural activity from multiple interacting brain regions.[4,46,58,71–73] Common methods to study interactions between brain regions such as linear regression,[37,40] CCA,[44] constrained dimensionality reduction,[38,39] linear discriminant analysis,[75] generalized linear models (GLMs),[37] or Granger causality[76] rely on correlative analysis of neural data, posing a few challenges. First, purely correlation-based inference of effective connectivity cannot distinguish between correlations that arise from common inputs and those that arise from other types of interactions between regions, though these can be partially accounted for by incorporating additional covariates.[37,77] Second, correlation alone does not provide directionality, though careful assessment of spike latencies can provide some insight into possible directional effects.[78] Third, correlative analyses typically describe interactions between two regions and are difficult to extend to data from multiple interacting regions, though recent work on switching dynamical systems shows promise.[79,80] CURBD addresses these limitations by building and analyzing RNNs that are trained to match the entire time-series from experimentally collected data[2,18] in a largely unbiased manner.

CURBD explicitly models the recurrence between all recorded neurons, capturing all possible multi-region interactions in the dataset. This allows us to, in an unbiased way, capture the directionality and magnitude of the interactions within and across regions that are responsible for the observed neural dynamics. Furthermore, the directed interaction matrix inferred from the trained multi-region RNN is not necessarily symmetric, allowing directional estimates of the inferred functional interactions (e.g., **Fig. 5e**). Crucially, the dynamical system is capable of generating activity patterns through a dynamical system consistent with the data, we capture more than the covariance between neuronal outputs. In other words, the covariation in neuronal outputs–and therefore, consistency with the "on-manifold" directions–is an inevitable consequence of our approach. We go beyond that however, due to the fact that we capture the on- and off-manifold directions. Therefore, the biologically relevant prediction we make is that our data-constrained RNN-based inference technique should be more robust, at least in part, to perturbations also in numerous off-manifold directions. In contrast to dynamic causal modeling,[81,82] CURBD does not necessarily require known perturbations or inputs, and can flexibly model any dataset. Lastly, since CURBD concurrently models all multi-region interactions, it scales natively to arbitrarily large datasets with any number of regions, even to whole-brain recordings, now available from *c. elegans*,[83] fruit fly,[84] and larval zebrafish.[4,7,71] Subsequent extensions of our approach could elucidate convergent mechanisms across species,[20,85] and identify unique divergences.[86]

Interpreting directed interactions inferred from RNNs constrained directly by data

While CURBD estimates multi-region interactions by incorporating recurrence within and between regions responsible for the observed neural dynamics, these interactions should not be considered causal relationships. Additionally, the directed interactions estimated by the Model RNN need not relate to actual synaptic connectivity. While a direct monosynaptic connection between two neurons should contribute to a strong directed interaction weight, strong interactions could arise indirectly as well.[87]. Polysynaptic pathways (including those involving neurons or brain regions that were not experimentally observed) or triggering neuromodulator release could enable one neuron to exert an influence on other neurons that would manifest as an inferred directed interaction weight.[7,88] Additionally, brain-wide state changes that impact distributed neural circuits—such as those induced by stress,[7] depression,[89] or even glia[90]—could lead to strong functional relationships between recorded neurons.

Model RNNs underlying CURBD are a specific dynamical system

The Model RNNs used for CURBD are specific, learned dynamical systems that capture the essential features of the neural dynamics from the data they were trained to match based on an initial condition. This facet represents a difference between CURBD and other approaches that seek generative models of the neural dynamics.[34] However, even though our models here are typically fit only to single instantiations of data,[91] we identify consistent solutions from one iteration to another, for instance, at the level of statistical distributions of groups of interaction weights (e.g., **Fig. S7**) as well as at the level of currents inferred by CURBD (e.g., **Fig. 5f**). Furthermore, since the inter-region currents inferred by CURBD rely on the product of the directed interaction weight matrix and the activity, the estimation noise in different realizations of the matrix are averaged out. Therefore, the currents identified by CURBD are much more robust to different random initializations of the directed interaction matrix, allowing for consistent solutions under a variety of initialization conditions, as well as to different random subsamples of the modeled neurons.

Ultimately, the Model RNNs underlying CURBD should be considered as a model of the data itself—an *in silico* representation of the experiment. This model enables a deeper dive into the experimentally measured data using the directed interaction matrix or currents due to inter-region interactions which we cannot access experimentally.[24] Our current approach assumes that a single directed interaction matrix captures the dynamics for the whole duration of the data. Factors such as learning[8,92] or behavioral state changes[7] could change the dynamical rules governing the interactions among different neural populations in vivo. If such state changes are sought, they can be identified by fitting different Model RNNs on different samples of data (e.g. periods of time, task conditions). The final currents can then be fully reconstructed by essentially "stitching together" the currents inferred by Model RNNs fit to each set of samples. More elegantly, the training process could also be modified, in future extensions of this work, to identify state changes and adjust the directed interaction matrix over time in a partially unsupervised, adaptive manner consistent with biological plasticity mechanisms.

Additional uses and extensions of CURBD

The multi-region Model RNNs employed in the applications above made no assumptions about the structure of the directed interaction matrix or inter-region connectivity. Instead, we allowed the neural networks to opportunistically, through the process of training, construct solutions that recapitulated the essential dynamical features in the multi-region experimental data. In biological systems, there are numerous anatomical constraints that could be incorporated into the model in the future. For example, the effect of a given neuron on its numerous downstream targets is sometimes thought to be either excitatory or inhibitory.[93] This constraint could be incorporated into the learning rule such that columns of the directed interaction matrix are restricted to have either all positive or all negative weights. Additionally, while we allow our RNNs to be weighted all-to-all, including the inter-regional interactions (the off-diagonal submatrices), inter-regional connections in biological brains are highly structured. For example, long-range connections between regions are likely more sparse than within a local population.[94] Such sparsity could be induced in an unsupervised manner by applying an L1 norm on the weights of specific inter-region submatrices in the cost function or by extending our previously developed "PINning" approach[6] for restricting the fraction of recurrent weights to be trained based on the data, to the case of inter-region submatrices. Brain-wide connectomics data[87,95,96] is becoming increasingly available and could also be leveraged to build structure a priori into the directed interaction matrix about which pathways should be directly connected. Future extensions of this work could focus on developing faster, online training and inference pipelines to see how well the directed interaction matrices inferred from data-constrained multi-region RNNs are either constrained by or predictive of the respective structural connectomes. Lastly, we trained the Model RNNs using rates estimated from the neural recordings. Future extensions of CURBD could allow more temporally-precise directed interaction estimates by incorporating models of spiking statistics into the model.[77,97]

In the present work, the region identity of each experimental neuron was known using anatomical landmarks or electrode implantation site.[7–10] This knowledge allowed us to readily divide the directed interaction matrix into region-specific blocks. However, we predict that in future work CURBD can be extended to provide a basis for functional clustering that goes beyond anatomical designations by applying clustering or tensor decomposition methods[98] directly to the currents inferred by the Model RNN into each target unit. CURBD could then be used in an unsupervised manner to find relevant population designations based on functional distinctions and their interactions with other neurons. This could identify interdigitated submodules within single regions[99–101] as well as identify brain-wide functional circuits[102], as we have illustrated here.

We used activity from single, identifiable neurons to constrain the Model RNNs for CURBD, but the possible use cases of the general approach are not confined to cellular resolution data. Model RNNs can be fit to non-cellular resolution data, such as multi-electrode local field potential recordings, which can serve as macroscopic estimates of similar population dynamics.[103] Furthermore, other types of relevant experimental data or conditions can be incorporated as additional constraints on the Model RNNs during training. Behavioral data such as body posture derived from modern pose detection methods[104,105] could be incorporated into the training process to help account for unobserved common inputs related to that behavior.[106] Static labels representing experimental metadata (behavioral task, stimulus condition, etc.) could also be incorporated to help compensate for brain-wide state changes. These measurable external signals could be targeted to all recorded neurons, or a specific subset (e.g. brain region, cell type) if such constraints are known. Importantly, all of the extensions described above do not change the fundamental principles underlying CURBD, which at its core, relies on straightforward and interpretable matrix multiplication. They only serve to provide a more constrained estimate of the biological system's directed interactions.[107]

The power of CURBD lies in harnessing the ability to flexibly engineer multi-region RNNs based on a broad range of time-series data from various experiments, as we have exemplified with the four applications presented here. There is often a remarkable conservation of structure and function throughout evolution and across species producing a certain behavior even with divergent phylogenetic trees.[85] Therefore, understanding the commonalities (or unique differences) in identified mechanisms across different species will be critical to uncover fundamental principles of neural computation.[11,20] This requires an analytical framework such as CURBD that robustly and flexibly scales across a range of different experimental factors—e.g., methodologies, levels of granularity, sampling densities, and spatiotemporal resolutions—such as those encountered when comparing different species ranging from smaller, highly sampled nervous systems (e.g., *Caenorhabditis elegans*, *Drosophila*, larval zebrafish) to larger, less sampled brains (e.g., rodents, non-human primates, and humans) (**Fig.**

2a). Thus, CURBD provides a powerful new approach for comparative studies over time, across individuals, across scales of neural function, or even across species.

Acknowledgements. M.G.P. is supported by the Fonds de recherche du Québec - Santé (chercheurs-boursiers en intelligence artificielle). C.A. is supported by a Louis Perry Jones, an Alice and Joseph Brooks, and a Mahoney Postdoctoral Fellowship. S.S. is supported by an European Molecular Biology Organization (EMBO) Postdoctoral Fellowship. M.B. is supported by the Gatsby Charitable Foundation GAT3708 and NIH R01EB029858. U.R. is supported by the National Institutes of Health (NIH) (R01 MH110831 and U01 NS117839) and National Science Foundation (NSF) (BCS-1554105). P.R. is supported by a National Institute of Mental Health BRAINS award (R01MH110822, R01MH118638) and a young investigator grant from the Brain and Behavior Foundation (NARSAD). C.D.H. is supported by an NIH Director's Pioneer Award (DP1 MH125776) and NINDS R01 NS089521. K.R. is supported by NIH BRAIN Initiative (R01 EB028166), James S. McDonnell Foundation's Understanding Human Cognition Scholar Award, and NSF FOUNDATIONS award (NSF1926800). K.D., T.B., and A.S.A. were supported by a TR01 gant from the NIH. We would like to thank Dr. Juan A. Gallego, Dr. Raeed H. Chowdhury, Dr. L. F. Abbott, Dr. Sheila Cherry, and Dr. Christian D. Marton for comments on earlier drafts of this manuscript. We are indebted to and inspired by Dr. C. R. Rajan and Prabha Rajan for sharing their undying love of learning.

Author contributions. M.G.P. and K.R. conceived of the method. M.G.P. analyzed datasets and generated figures. M.G.P., M.B., K.D., and K.R. wrote the manuscript. M.G.P., M.B., C.A., S.S., S.Z., M.B., C.M., J.M., U.R., P.R., K.D., C.D.H., K.D., and K.R. edited the manuscript. E.C. ran cloud simulations and streamlined prototyped code. A.S.A., T. B., and K.D. provided the larval zebrafish dataset. C.A., S.S., and C.D.H. collected and processed the mouse dataset. M.E.Y., C.M., and P.R. provided the macaque dataset. J.M. and U.R. provided the human dataset. K.D. and K.R. supervised all aspects of this work.

Conflict of interest statement. The authors declare no competing interests.

Data availability statement. The human datasets are publicly available at: . The remaining fish, mouse, and monkey datasets can be shared upon reasonable request.

Code availability statement. All modeling and analysis in this manuscript was done in Matlab (The Mathworks, Inc.) or Python. Matlab and Python code to train multi-region Model RNNs based on experimental recordings and perform CURBD using the inferred interactions is available at: https://github.com/rajanlab/CURBD.

Methods

Multi-region recurrent neural networks

Network elements. Network models represent real biological circuits, but they do so with different levels of fidelity.[19,22,24] We constructed Model RNNs that are directly constrained by experimentally-obtained time-series neural data. Each network unit or model neuron, here indexed by i is described by a total current $x_i(t)$, and $r_i(t)$, a nonlinear function ϕ of $x_i(t)$, where $i=1$, 2, …, N is the total number of units in the network. Each variable $x_i(t)$ obeys the following equation:

$$\tau \frac{dx_i(t)}{dt} = -x_i(t) + \sum_j^N J_{ij} r_j(t) + h_i(t) \tag{4}$$

where h_i is the external input to the unit, **J** is a heterogeneous matrix of recurrent connections, and τ is the unit's time constant selected based on the data to be modeled (for the datasets in this manuscript, see **Table 2** below). The sum of $r_j(t)$

weighted by J_{ij} represents the current received by unit i from all network units, which will be filtered by the leaky integration and give rise to $x_i(t)$, the current going into unit i. We use ϕ=tanh(x), but other saturating nonlinearities, such as sigmoids, have been explored in related work (see Refs. [6,42]).

In Equation 4, we use parentheses $x(t)$ to represent that the variable is continuous in time. In practice, the network equations are integrated using the Euler method and an integration time step, Δt_{RNN}. Note that we allow network integration to interpolate with a finer time step than the sampling rate of the experimental data to be modeled. This allows for smoother dynamics when the experimental data may be relatively sparsely sampled. In such discrete implementations, we use square brackets $x[t]$ to represent that the time is defined in a discrete domain:

$$x_i[t] = x_i[t-1] + \tfrac{\Delta t}{\tau}\left(-x_i[t-1] + \textstyle\sum_j^N J_{ij} r_j[t-1] + h_i[t-1]\right) \tag{5}$$

Recurrent weights carrying inputs onto a target unit $i=1$, 2, …, N from its source partner j, J_{ij}, which are the elements of the matrix **J**, are either fixed or modifiable (plastic), depending on how much structure is introduced into the connectivity of the initially disordered network. We introduce no *a priori* structure in **J**, allowing all elements to be modified during training. J_{ij} can potentially be modified by a number of different learning algorithms (see below). The **J** matrix is arranged such that each block along the diagonal represents the recurrent connections within each brain region being considered, and the off-diagonal blocks, the inter-region projections to and from them. In this way a two-region Model RNN has two blocks on the main diagonal relating each region to itself, and two regions on the off-diagonal relating each region to the other (e.g., **Fig. S1f**). Following this pattern, multi-region Model RNNs can be initialized with **J** matrices containing more than 2 blocks (e.g., **Fig. 5e**).

Typically, the initial, untrained directed interaction matrix $\mathbf{J}_0$ is constructed to be the same size as the number of neurons in the dataset to be modeled, though larger networks in which different subsets of weights are modified in a data-dependent manner have been explored previously in Ref. [6]. Here, each Model RNN unit is matched to one recorded neuron from the respective experimental dataset. The individual weights in $\mathbf{J}_0$ are initially chosen independently and randomly from a Gaussian distribution with mean and variance given by $<\mathbf{J}_0>=0$ and $<\mathbf{J}_0^{\,2}>_J=g^2/N$. The control parameter g determines the strength of the recurrent connections, and thus whether (g>1) or not (g<1) the network produces spontaneous activity with non-trivial dynamics.[23,42,52] We set g=1.5 here, though in practice we observe qualitatively similar results for a range of values provided g is sufficiently large to facilitate chaotic spontaneous dynamics in the network in the absence of inputs. Ultimately, the elements of **J** will be modified by the training algorithm until the activity of the model RNN's units match the target neural data, and after training, autonomously produce dynamics consistent with the experimental recordings.

Design of external inputs. In biological neural circuits, populations of cells are constantly driven by external and inter-regional inputs that we cannot always observe. To mimic this effect without modeling entire sensory-motor transforms and pathways, and to keep the external drive as general as possible, the external inputs to the each unit i in the Model RNN, denoted by $h_i(t)$, are generated from filtered and spatially delocalized white noise that is frozen, using the equation:

$$\tau_{WN}\frac{dh_i}{dt} = -h_i(t) + h^0\eta(t) \tag{6}$$

where η is a random variable drawn from a Gaussian distribution with 0 mean and unit variance, and the parameters h^0 and τ_{WN} control the scale of these inputs and their correlation time, respectively. We use h^0=1 and τ_{WN}=0.1 in this paper. There are typically as many different inputs as there are model neurons in the network, with individual model neurons receiving the same input on every simulated trial in examples where there is an explicit trial structure (such as the monkey and human datasets used in this paper).

Model RNN training. The directed interaction weight matrix **J** is trained by a target-based method in order to match the network activity with the experimental data from cellular-resolution neural recordings or imaging. As the target is discrete in time, we take a discrete implementation of the RNN dynamics as in Equation 5, where we use square brackets to note the discrete domain. During training, the activity of individual units in the Model RNN is compared directly to target functions derived from the experimentally-recorded neurons, denoted by $a_i[t]$, to yield error values $e_i[t]$ that will be used to update **J** The error can be computed based on the influx current $x_i[t]$ ("current-based" training) or the firing rate $r_i[t]=\phi(x_i[t])$ ("rate-based" training) of individual units. .

For "current-based" training, the target functions $a_i[t]$ are for the currents into each RNN unit. If the original experimental data is firing rates, it would need to be translated into current-like quantities by inverting the chosen activation function. At each time step, the influx current of individual network units, $x_i[t]$, is treated as though it were being evaluated at steady state based on Equation 5, and then compared with the target functions $a_i[t]$ to yield the error value $e_i[t]$:

$$e_i[t] = x_i[t] - a_i[t] \tag{7}$$

The error function is a linear function of the parameter being optimized (**J**), so we can adopt a learning rule similar to Recursive Least Squares (RLS) approaches.[6,25,108] During training, the elements in the directed interaction matrix **J**, J_{ij}, undergo modification at a rate proportional to three factors: i) the error term computed above; and a regularization term consisting of, ii) the "presynaptic" or source firing rate of each neuron; and iii) a matrix **P** with N x N elements (defined below in Equation 10).

Training proceeds iteratively as schematized in **Fig. 2b**. At each learning time step t (fixed to a multiple of the integration timestep in the RNN dynamics) for $i=1, 2,..., N$ target units, the corresponding elements of **J** are adjusted from their values at the previous time step ($t-1$), as introduced in Equation 1, according to:

$$J_{ij}[t] = J_{ij}[t-1] - \Delta J_{ij}[t] \tag{8}$$

where the update term[25,108] is computed according to:

$$\Delta J_{ij}[t] = e_i[t] \sum_{k}^{N} P_{jk}[t] r_k[t-1] \tag{9}$$

In the above, **P** is defined mathematically as the inverse cross-correlation matrix of the firing rates of units in the network, such that its elements P_{ij} are given by:

$$P_{ij} = \left(\mathbf{C}^{-1}\right)_{ij}, \text{ where } C_{ij} = \langle r_i r_j \rangle \tag{10}$$

The matrix **P** tracks the correlations in the firing rate fluctuations across the whole network at every time step, and is computed for all $i=1$, 2,..., N target units and all $j=1$, 2, ..., N source units. It is not generally necessary to calculate the matrix **P** explicitly. Instead, **P** can be updated iteratively[108] according to (see **Fig. S11** for details).

$$\mathbf{P}[t] = \mathbf{P}[t-1] - \frac{\mathbf{P}[t-1]\mathbf{r}[t-1]\mathbf{r}'[t-1]\mathbf{P}[t-1]}{1 + \mathbf{r}'[t-1]\mathbf{P}[t-1]\mathbf{r}[t-1]} \tag{11}$$

where **r**[*t*] denotes a vector whose *i*th element is $r_i[t]$. The matrix **P** is initialized to the identity matrix scaled by a factor P^0 which controls the overall learning rate. In practice, training is most effective when P^0 is set to be 1 to 10 times the overall amplitude of the external inputs (h^0).

As an alternative to the "current-based" training, the network can be trained using a "rate-based" learning rule, The target function $a_i[t]$ is now for the firing rate of units. The RNN unit firing rates, $r_i[t]$, are directly compared to the targets to yield the error values:

$$e_i[t] = r_i[t] - a_i[t] \tag{12}$$

The rate-based training is similar to the current-based training and takes the same updating rule as in Equation 9, if we replace $e_i[t]$ with the rate-based definition. Therefore, the choice of learning rule (current-based or rate-based as described above) would impact the calculation of Δ**J**.

For the interpretation of the rate-based training, the update term Δ**J** to the matrix in this case amounts to a "gradient-like" approach, albeit one with a multiplicative regularizer mediated by **P**. This multiplicative term makes a summary of the activation of all the units, weighted (regularized) by **P**, which potentially balances the global activation. This is not present in standard gradient-based training for RNNs. The key distinction between the rate-based and current-base methods is that in rate-based training, we do not need to make any assumptions about the nonlinearity of the biological neuron that we are targeting, while in current-based training such a nonlinearity would need to be explicitly inverted to produce the current-like target functions from rate data. Additional implications of this type of learning rule and a rigorous comparison between the two formalisms, beyond **Fig. S2** here, is for future work.

Since the learning algorithm updates **J** at each learning time step, accurate fits could be observed during training even when the algorithm has not fully converged. Thus, after training for a fixed number of iterations (typically between 1500 and 3000 iterations for model RNNs based on experimental neural datasets), we stopped the training to run the "learned" autonomous dynamical system for a few additional iterations to compute and evaluate the final goodness of fit. We assessed the quality of the fit and convergence using two metrics: 1) the training error ($\boldsymbol{\chi}^2$) between the Model RNN rates and the teacher/target functions (derived from data in the current-based training algorithm, or the time-series directly, in the case of rate-based training algorithm), computed as the mean-squared error $e_i(t)$ along all $i=1$, 2,…, N target neurons; and 2) the proportion of variance explained (pVar) as one minus the ratio of the average squared difference between the neural data and outputs of the network compared to the variance of the data:

$$\mathrm{pVar} = 1 - \frac{\langle e_i[t]^2 \rangle}{\langle (a_i[t] - \bar{a}[t])^2 \rangle} \tag{13}$$

where $\bar{a}[t]$ is the mean across all units at each time step and the triangle brackets indicate average across time points and neurons.

Analyzing the Directed Interaction matrix J after training. The directed interaction matrix inferred by the Model RNN quantifies the strength of interactions between the units in the network. These values can be either positive or negative, suggesting excitatory or inhibitory effects on the target neuron, respectively. Since the RNNs we build are extensively constrained by neural dynamics, we find that it is possible to consistently infer similarly distributed matrices, when starting from different random realizations of the initial weight distribution (e.g., **Fig. S7**). Thus the statistical properties of the interaction strengths we derive from data-constrained RNNs can be reliably compared across brain regions, as well as between RNNs trained to match a range of experimental datasets from different species. In this paper, we summarized the statistical properties of such model-derived interaction strengths by computing histograms using the total number of elements of either the full **J** matrix or specific submatrices containing the strength and type of interactions within and between individual brain regions. Notably, when analyzing these matrices, we scaled the distributions by the square root of the number of source units to account for differences in population sizes. We also normalized each histogram by the maximum value to facilitate comparison between matrices derived from RNNs of different sizes, and visualized the distributions using a logarithmic scale. These distributions could be further summarized and quantified by metrics such as the median, standard deviation, skewness, or kurtosis.

Computing current sources to specific brain regions. As indicated by Equation 4, the time derivative of the net current going into one unit in the Model RNN, $dx_i(t)/dt$, is proportional to the current it receives from all source units, defined as the product of the corresponding row of the directed interaction matrix and the activity of all source units at the previous time step. We define the total input current the i^{th} target unit received as $I_i(t)$, which is a linear combination of the rates of all the (source) units in the networks::

$$I_i(t) = J_{i1}r_1(t) + J_{i2}r_2(t) + ... + J_{iN}r_N(t) \tag{15}$$

CURBD adopts this linear decomposition to study brain-wide currents between active neurons across multiple interacting brain regions. In this manuscript, we computed the currents in the target regions using the weights in different submatrices as described here. However, this method can be readily extended to separately infer excitatory (or inhibitory) currents by first setting all of the negative (or positive) values in the **J** matrix to be zero and then repeating the summation in Equation 15.

Due to the large number of free parameters in the Model RNN, i.e., order N^2 elements for RNNs with N units, the training algorithm does not necessarily infer the precise entries, element-by-element, in the directed interaction matrix, even when ground truth simulated data originated via low-rank or smoothed connectivity (for details, see Ref. [6]). However, we find consistent and reliable estimates, i.e., recapitulating statistical properties of groups of weights in the directed interaction matrices. Furthermore, after training, the RNNs are able to produce highly consistent dynamics even when starting from different initial conditions. In practice, when taking the dot product of **J** and **r**(t) to compute the currents for CURBD, random element-by-element fluctuations in the individual reconstructed weights between pairs of units are averaged out, but the overall population dynamics are preserved. For this reason, in its current state, CURBD is best applied to infer interactions between source and target brain regions with sufficient numbers of active neurons. Future extensions, e.g., those that incorporate known connectivity between regions[87,96] or additional constraints from data, such as behavioral covariates, could provide reliable current estimates with finer granularity than at the level of individual regions and possibly across different behavioral states.

Two-region model producing idealized, ground truth, simulated data to validate CURBD

Design of the generator model. We simulated a model that generated idealized ground truth data to test when CURBD approach would be the most effective at disentangling inter-region interactions, and to probe the conditions under which it would fail to perform optimally. We generated two 1000 unit RNNs, each with random connectivity weights drawn from a Gaussian distribution, as described in the initialization procedure for the Model RNN above. One RNN (corresponding to Region A) was driven by an external sinusoidal signal, $S_A[t]$, oscillating at $\pi/8$ Hz, while the second (corresponding to Region B) was driven by another sinusoid, $S_B[t]$, oscillating at $\pi/3$ Hz and phase shifted by $\pi/3$. The two sinusoidal inputs began after two seconds of a simulated “resting state” during which the inputs to the RNNs were set to zero. These external inputs were connected to 33% of the units in their respective RNN with a fixed input weight, picked from a uniform distribution. The two RNNs were recurrently connected, with a varying percentage of neurons in each region (randomly selected) receiving inputs from the other region with a fixed weight of one. We computed the time-series current activity of the i^{th} unit in the two RNNs, $x_{A,\,i}[t]$ *and* $x_{B,\,i}[t]$ for ten seconds of data using the following steps. We initialized the states of the two RNNs to random values between -1 and 1. For each subsequent time step, we computed the change in activity of each RNN unit *i* based on its inputs according to:

$$\Delta x_{A,i}[t] = \mathbf{J}_A r_{A,i}[t-1] + w_{rgn} C_{BtoA,i} r_B[t-1] + w_{in} C_{StoA,i} S_A[t-1] \tag{16}$$

$$\Delta x_{B,i}[t] = \mathbf{J}_B r_{B,i}[t-1] + w_{rgn} C_{AtoB,i} r_A[t-1] + w_{in} C_{StoB,i} S_B[t-1] \tag{17}$$

C_{StoA} and C_{StoB} define binary connectivity vectors describing the connectivity of the external sinusoidal inputs to their respective regions. Similarly, C_{AtoB} and C_{BtoA} are binary vectors describing the connectivity between regions A and B. The fraction of entriess in the above inter-region connectivity vectors set to 1 is defined as p_{rgn}. w_{rgn} and w_{in} are scalars that set the connection weights for the sinusoidal inputs and inter-region connections, respectively. The $\mathbf{J}_A$ and $\mathbf{J}_B$ matrices are initialized to g_A^2/N and g_B^2/N, where the scaling parameters g_A and g_B control how chaotic each RNN is, as described in the Model RNN training section above. The activity of each RNN unit *i* was then computed according to:

$$x_{A,i}[t] = x_{A,i}[t-1] + \frac{\Delta t\left(-x_{A,i}[t-1] + \Delta x_{A,i}[t]\right)}{\tau} \tag{18}$$

$$x_{B,i}[t] = x_{B,i}[t-1] + \frac{\Delta t\left(-x_{B,i}[t-1] + \Delta x_{B,i}[t]\right)}{\tau} \tag{19}$$

Lastly, the activity of each RNN unit *i* was transformed into a firing rate by passing through the nonlinearity, as described above:

$$r_{A,i}[t] = tanh(x_{A,i}[t]) \tag{20}$$

$$r_{B,i}[t] = tanh(x_{B,i}[t]) \tag{21}$$

Checking robustness of CURBD over a range of simulation parameters for the Generator model. We repeated the ground truth simulations sweeping over a broad range of parameters applicable to the generator model (**Fig. S1k**): 1) g_A, the dynamical regime of Region A; 2) w_{rgn}, the strength of the weights of the recurrent connections between the networks; and 3) p_{rgn}, the proportion of neurons in Regions A and B receiving input from the other region. This parameter-sweeping process helped us explore how effectively CURBD operates to untangle currents resulting from the external sinusoidal inputs as the properties of the modeled networks change. The remaining parameters were fixed for all of these simulations. Values for all of the parameters we tested are provided in Table 1.

For each combination of parameters in the generator model, we trained a 2000-unit Model RNN to reproduce the activity of the two Regions from the generator model using the algorithm described above. The parameters chosen for this "fit" Model RNN (Table 2) were held fixed to ensure we studied the effect *only* of the network properties as we swept parameters, not of variations in the Model RNN. For each of these Model RNNs, we applied CURBD to Region A to assess how effectively we could isolate the currents from the two external inputs (the sinusoidal input driving Region A and the sinusoidal input driving Region B) from the population. Since each external input was effectively one-dimensional, we first reduced each estimated population-wide source current to a single component using PCA. We then computed how accurately we could infer the external inputs by directly comparing the correlation coefficient (denoted by R^2 in **Fig. S1j**) between the leading PC of each source current and each of the two sinusoidal external inputs. We repeated this analysis for different combinations of parameters to assess which parameter regimes consistently gave the highest R^2 values.

Three-region ground truth simulation to validate CURBD

Design of the generator model. We designed a second idealized ground truth simulation to validate whether CURBD could effectively infer source currents between different interacting regions even when there are no external inputs driving a particular neural population. We simulated three 1000-unit RNNs using a generator model similar to the two-region model described above (**Fig. 2d**). However, rather than sinusoidal inputs as in the two-region ground truth model, two of the three interconnected RNNs received time-varying patterns of inputs from other model networks. The external inputs driving Region B were provided by a network generating a Gaussian "bump" propagating across the network sequentially, *SQ*(*t*). The sequence began 2 seconds after the start of the simulation and ended 4 seconds later, with each sequentially activating unit *i=1*, 2,…, *N* behaving according to:

$$SQ_i(t) = e^{-\frac{1}{2\sigma^2}(\frac{i-\sigma-Nt}{T})^2} \tag{22}$$

where σ denotes the width of the bump across the population (here, 20% of the units), *N* represents the population size (here, 1000 units), and *T* represents the total simulation time of twelve seconds. The external input to Region C was provided by another 1000-unit network generating a fixed point, *FP*(*t*), for 8 seconds that instantaneously shifted to a new fixed point for an additional 4 seconds. The fixed points were generated by sampling *SQ*(*t*) at two different time points (*t*=2s and *t*=5s) and holding them at the sampled value of firing rate for the duration of the fixed point. The external inputs were connected to 50% of the units in their respective regions (randomly selected) with a fixed negative weight (inhibitory) for Region B and positive weight (excitatory) for Region C. The third RNN (Region A) received only the recurrent inputs from the other two RNNs, no external drive. The Region A RNN was modeled at a different value of *g* from the other two networks, yielding distinct dynamics (Table 3). The following update equations (again, discrete implementations with square brackets) governed the interactions of the regions at each time step (with a resolution of *Δt=0.01*), with the subsequent activity evolving similarly to the two-region simulation described by Equations 18-21:

$$\Delta x_{A,i}[t] = \mathbf{J}_A r_{A,i}[t-1] + w_{rgn} C_{BtoA,i} r_B[t-1] + w_{rgn} C_{CtoA,i} r_C[t-1] \tag{23}$$

$$\Delta x_{B,i}[t] = \mathbf{J}_B r_{B,i}[t-1] + W_{rgn} C_{AtoB,i} r_A[t-1] + w_{rgn} C_{CtoB,i} r_C[t-1] + w_{in} C_{SQtoB,i} SQ[t-1] \tag{24}$$

$$\Delta x_{C,i}[t] = \mathbf{J}_C r_{C,i}[t-1] + W_{rgn} C_{AtoC,i} r_A[t-1] + w_{rgn} C_{BtoC,i} r_B[t-1] + w_{in} C_{FPtoC,i} FP[t-1] \tag{25}$$

As in the previous simulation, the **J** matrices for each of the three networks were initialized based on the selected g_{A}, g_{B}, and g_{C} parameters.

Description of the Model RNN and CURBD analysis. We trained a 3000-unit Model RNN to match the activity of the three-region generator model using the procedure described above. We found that the Model RNN reproduced the simulated data accurately over a wide range of parameters; for the simulations reported in this paper (**Fig. 2 and S1**), we used the values reported in Table 2. We performed CURBD to infer the nine source currents governing interactions between the three regions. We reduced the dimensionality of the full 1000-unit population of each of the three regions and the source currents using PCA. We chose the leading five dimensions for the following analyses, which sufficed to capture more than 95% of the total variance in each source current, though we observed similar results with other assumed dimensionalities (data not shown).

Since the true connectivity of the network was defined in the simulated dataset, we computed the ground truth currents between each region to isolate the effect of isolated inputs from the source regions on the target region, including how the input activity would propagate through the recurrent connections of the target region. We adapted the update equations defined above to compute the three current sources into one region at each time step, here using Region A as an example:

$$Ix_{GT,AtoA,i}\,[t] = \mathbf{J}_A r_{A,i}\,[t-1] \tag{26}$$

$$Ix_{GT,BtoA,i}\left[t\right] = \mathbf{J}_A W_{rgn} C_{BtoA,i} r_B \left[t-1\right] \tag{27}$$

$$Ix_{GT,CtoA,i}\left[t\right] = \mathbf{J}_A W_{rgn} C_{CtoA,i} r_C \left[t-1\right] \tag{28}$$

The same process was performed for Regions B and C using similar equations. We performed the same dimensionality reduction analysis on the ground truth currents as on the inferred source currents from CURBD. Since the Model RNN was trained to reproduce time-varying activity from all the units in the multi-region generator model, each inferred source current has the same dimensionality as the ground truth current, and is embedded within the same high-dimensional space of the population activity of the respective simulated region. We could thus directly compare each leading PC using VAF (**Fig. 2j and S1**).

$$\mathrm{VAF} = 1 - \frac{\sum_t^T \left[Ix_{GT}(t) - Ix(t)\right]^2}{\sum_t^T \left[Ix_{GT}(t) - \bar{I}x_{GT}(t)\right]^2} \tag{29}$$

Comparison of inferred inter-region currents to shared dynamics identified by Canonical Correlation Analysis. We compared the performance of CURBD to an analogous decomposition obtained by canonical correlation analysis (CCA).[43,109] CCA obtains an optimal linear transformation relating the dimensionality-reduced population activity of the source and target regions to identify shared dynamics. In brief, we first took the low-dimensional trajectories of each region and performed a QR decomposition to identify for each region of the resulting **Q**, which provides an orthonormal basis for the column space of the low dimensional trajectories. For any pair of regions, for example Region A and Region B, we performed a singular value decomposition of the inner product of the corresponding **Q** matrices:

$$\mathbf{Q}_A' \mathbf{Q}_B = \mathbf{USV}' \tag{30}$$

This process effectively finds new dimensions within the manifold of Region A (denoted by **U**) and Region B (denoted by **V**) that maximize the correlation between the two trajectories. To analyze the shared dynamics between the two regions, we projected the activity of either region onto the corresponding axes. Unlike CURBD, the mapping obtained from CCA (and similar methods of inferring functional connectivity only from the covariance matrix of recorded neural activity) is not directional and is purely correlational. Thus, only one "current" can be obtained for each pair of regions. We compared the VAF by the first component identified by CCA to the first PC of the ground truth currents to assess the effectiveness of this approach (**Fig. S1c,d**).

Addressing partial sampling issues present in experimental data. We repeated the above simulation to determine whether or not CURBD is effective when only a fraction of the total multi-regional activity is 'observed' by the Model RNN. This control analysis addresses partial sampling issues present in real data when activity can be experimentally measured from only a relatively small fraction of the total number of neurons in a region. To simulate this scenario and test the efficacy of CURBD in the face of partial sampling issues, we trained Model RNNs to match activity from 5%, 20%, 50%, and 100% of the available neurons in each region (randomly selected) of the ground truth multi-region generator model. We repeated the simulation ten times at each subsampling level to help account for variability in the random sampling of neurons, as well as variability from different random initializations of the **J** matrix. Such variability scales inversely with network size for Gaussian weights; the ten repetitions at 100% sampling thus provide a lower-bound on the variability that would be expected within this model. Unlike the initial Model RNN analysis where every neuron was sampled, in the subsampling case, we can no longer guarantee that the axes should be oriented similarly in PC space and VAF is not a reliable measure of how well the method performed. Thus, here we again employed CCA not to identify shared dynamics between regions, but to compensate for differences in the number of sampled neurons generating the dynamics of a single region,[43] In this application, CCA provides a quantitative “similarity index”–quantified by the canonical correlation of the leading aligned dimension–of the population dynamics between the currents identified by CURBD and the ground truth currents (**Fig. 2k**).

Multi-region calcium fluorescence recordings in larval zebrafish

Experimental setup. All procedures were approved by the Stanford University Institutional Animal Care and Use Committee. The experimental details have been previously described in Ref. [7]. In brief, five larval zebrafish expressing nucleus-localized GCaMP6s — homozygous Tg(elavl3:H2B-GCaMP6s)[54] — were immersed in water and partially head-fixed in agarose (Thermo Fisher Scientific 16250-100), leaving their tails free to attempt swimming movements. The fish was illuminated with 850nm infrared light and tail movements were recorded using a high-speed camera (Allied Vision Technology Manta G-031B) and macro lens (Nikon AF-S DX Micro Nikkor 85 mm f/3.5G ED VR) through both short-pass (Thorlabs FES0900-1) and long-pass (Thorlabs FEL0800-1) filters. The spontaneous neural activity and movements of the fish were recorded for an extended period of time (48+ minutes). In one cohort of 5 fish, gentle but inescapable electrical shocks were applied to the water in the dish[7] which induced stress responses. In a second cohort of 5 fish, no shocks were applied but the recordings were made for the same duration.

Image acquisition and processing. Images were acquired with an Olympus FVMPE multiphoton microscope (Olympus) using fast piezo axial scanning at 1-1.2 Hz to image a volume of 562 x 241 x 76mm on average with 6-8 mm between planes. 920nm light was used to excite the GCaMP6S sensors. Light emission was collected through a 485 nm dichroic mirror and 495-540 nm emission filter. After collection, drift between images was corrected using an affine robust alignment by sparse and low-rank decomposition[110] and regions of interest (ROIs) were identified as the 90th percentile of each plane. These ROIs were pruned based on size and template-matching to nucleus-sized disks to isolate putative neurons. Fluorescent activity (F) was normalized to compute a typical $\Delta F/F$ trace by subtracting and dividing by the fluorescence value of all pixels within each ROI and defining F as the 5th percentile of all values in the session. Cells were then manually assigned to brain regions based on anatomical landmarks. In this work, we considered only cells identified as part of the telencephalon, habenula, thalamus, or raphe nucleus (Table 4).

Model RNN fitting. We used the 2-region larval zebrafish dataset to compare CURBD output using both current-based and rate-based learning rules. We used the 3-region larval zebrafish dataset to study the mechanisms underlying behavioral state transitions. For each zebrafish, we trained Model RNNs with 2 regions (Fish cA: 4372 units; Fish cB: 2937 units; Fish cC: 1578 units; Fish cD: 4303 units; Fish cE: 2171 units) or 3 regions (Fish cA: 5432 units; Fish cB: 3861 units; Fish cC: 1678 units; Fish cD: 5063 units; Fish cE: 2333 units; Fish sA: 4704 units; Fish sB: 3908 units; Fish sC: 1784 units; Fish sD: 2177 units; Fish sE: 3857 units) to reproduce the time-series Ca^{2+} data from the multi-regional activity. We used identical model parameters for all fits to all fish (Table 2). Before training each model, we randomly initialized the **J** matrix and white noise inputs as described above. For the 2 region models, in order to most accurately compare the learning rules, we fixed these initial conditions and trained the Model RNN to fit the experimental data using one of the two learning rules. We then repeated this procedure five additional times to assess the reproducibility over runs. Each data sample contained 6 minutes of recordings from the control cohort of fish. For the 3-region models, we used rate-based training and divided the 48 minute dataset into 8 epochs of 6 minutes each. We sequentially trained Model RNNs on each epoch starting from an initial random **J** matrix in epoch 1, and using the resultant **J** post-training as the initial guess for the subsequent epoch (a process called *chaining*). We then concatenated the current estimates identified within each epoch to reconstruct the interactions in the full 48 minute dataset.

Analysis of model output to compare learning rules. We compared the similarity of the resulting **J** matrices after rate-based and current-based learning by computing the two-dimensional correlation of the matrices (**Fig. S2**). As a reference, we compared the correlations obtained across different training runs with the same learning rule as well as after shuffling each matrix. We then used CURBD to compute the current sources driving the telencephalon populations. We compared the currents obtained by each learning rule by using PCA to compute the leading 10 components and computing the vector norm at each time point within this reduced dimensional space (**Fig. S2i**). We lastly computed the similarity within this low-dimensional space using CCA as described above (**Fig. S2j**).

Analysis of inter-region currents during the behavioral challenge paradigm. We analyzed the behavior of the two cohorts of fish. We identified timestamps of each tail whipping movement using the high speed cameras monitoring the fish. We then convolved this sequence of timestamps with a Gaussian kernel (60s kernel width) to produce an estimate of the continual movement rate. We then averaged this movement rate across fish in each cohort (**Fig. 3i**). To study the large-scale temporal dynamics of each current source, we decomposed the regional inputs to the Habenula and then performed PCA separately to reduce the activity in the space of habenular neurons to a single leading dimension (**Fig. S4**). We then averaged this trajectory across fish in each cohort to compare the temporal modulation of the dominant patterns of inputs from each region to habenula (**Fig. 3j**). Lastly, we analyzed the directed interaction weights to look for possible underlying mechanisms for the current dynamics. In each epoch we computed the distribution of weights in the sub-matrix capturing inputs from the raphe nucleus to the habenula. We computed this distribution for each fish of the shocked cohort in the 6 minute baseline epoch (pre-shocks), the first epoch of shocks (active coping), and the final epoch of the experiment (passive coping) and averaged these distributions across fish (**Fig. 3k**).

Multi-region calcium fluorescence recordings in mice

Surgery. All experimental procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee and were performed in compliance with the Guide for the Care and Use of Laboratory Animals. Two female mice expressing GCaMP6s (C57BL/6J-Tg(Thy1-GCaMP6s)GP4.3, The Jackson Laboratory, stock 024275) were implanted with cranial windows over the cortical surface. Mice were 3-5 months old at the time of surgery, and given an injection of dexamethasone (3 µg per g body weight) 4-8 h before the surgery. Mice were anesthetized with isoflurane (1-2% in air). A cranial window surgery was performed to either fit a 'crystal skull' curved window (LabMaker UG) exposing the dorsal surface of both cortical hemispheres,[47] or to fit a stack of custom laser-cut quartz glass coverslips (three coverslips with #1 thickness each (Electron Microscopy Sciences), cut to a 'D'-shape with maximum dimensions of 5.5 mm medial-lateral and 7.7 mm anterior-posterior, and glued together with UV-curable optical adhesive (Norland Optics NOA 65), exposing the left cortical hemisphere. The dura was removed before sealing the window using dental cement (Parkell). A custom titanium headplate was affixed to the skull using dental cement mixed with carbon powder (Sigma-Aldrich) to prevent light contamination. A custom aluminum ring was affixed on top of the headplate using dental cement. During imaging, this ring interfaced with a black rubber balloon enclosing the microscope objective for light-shielding.

Imaging and behavior setup. Data were collected using a large field of view two-photon microscope assembled as described in Ref. [58]. In brief, the system contained a combination of a fast resonant scan mirror and several large galvanometric scan mirrors allowing for especially large scan angles. Paired with a remote focusing unit to rapidly move the focus depth, this setup enabled random access imaging in a field of view of 5-mm diameter with 1 mm depth. The setup was assembled on a vertically mounted breadboard whose XYZ positions and rotation were controlled electronically via a gantry system (Thorlabs). Thus, to position the imaging objective with regards to the mouse, the position and rotation of the entire microscope were adjusted while the position of the mouse remained fixed. Mice were head-fixed and placed on an air-suspended 8-inch diameter Styrofoam spherical treadmill that enabled spontaneous running. Using two optical sensors (ADNS-9800, Avago Technologies), we tracked the treadmill velocity, which was translated into pitch, roll, and yaw velocity using custom code on a Teensy microcontroller (PJCR) as a readout of the mouse's running speed and direction. Individual recording sessions lasted from 45–60 minutes. Mice were extensively acclimated to head-fixation and running on the treadmill before data collection. We recorded behavioral and neural activity while mice spontaneously ran on the ball. The room was kept in complete darkness throughout the experiment. We defined running bouts as periods when the ball movement speed crossed a fixed threshold set to be the 90th percentile of the running speeds throughout the session.

Image acquisition. The excitation wavelength was 920 nm, and the average power at the sample was 60-70 mW. The microscope was controlled by ScanImage 2016 (Vidrio Technologies). We targeted four distinct regions in the left cortical hemisphere: primary visual cortex (V1), secondary motor cortex (M2), posterior parietal cortex (PPC), and retrosplenial cortex (RSC). These regions were targeted based on retinotopic mapping (see below). In each region, we acquired images in layer 2/3 from two planes spaced 50 μm in depth, at 5.36 Hz per plane at a resolution of 512 x 512 pixels (600 μm x 600 μm).

Retinotopic mapping for selecting Ca2+ imaging locations. We performed retinotopic mapping in the mice used for calcium imaging experiments as previously described in Ref. [8]. Mice were lightly anesthetized with isoflurane (0.7–1.2% in air). GCaMP fluorescence was imaged using a tandem-lens macroscope where excitation light (455 nm LED, Thorlabs) was filtered (469 nm with 35 nm bandwidth, Thorlabs) and reflected onto the brain through a camera lens (NIKKOR AI-S FX 50 mm f/1.2, Nikon) focused 400 μm below the brain surface. GCaMP emission light was collected using the same lens, filtered (525 nm with 39 nm bandwidth, Thorlabs), and imaged with another camera lens (SY85MAE-N 85 mm F1.4, Samyang) and a CMOS camera at 60 Hz (ace acA1920-155um, Basler). These images were synchronized to visual stimuli presented on a gamma-corrected 27 inch IPS LCD monitor (MG279Q, Asus). The monitor was centered in front of the mouse's right eye at an angle of 30 degrees from the mouse's midline. The visual stimulus, a spherically corrected black and white checkered moving bar[111] (12.5 degree width, 10 deg/s speed), was presented in 7 blocks, each consisting of 10 repeats of 4 movement directions (up, down, forward, backward). To produce retinotopic maps, we calculated the temporal Fourier transform at each pixel of the imaging data and extracted the phase at the stimulus frequency.[112] These phase images were averaged across repetitions for a given movement direction and smoothed with a Gaussian filter (25 μm s.d.). Lastly, we calculated field sign maps by computing the sine of the angle between the gradients of the average horizontal and vertical retinotopic maps.

For each retinotopic mapping session, we also acquired an image of the superficial brain vasculature pattern under the same field of view. We then acquired a similar brain vasculature image under the large field of view two-photon microscope. These two reference images were manually aligned and used to directly locate V1 and PPC locations for two-photon imaging. The location for RSC imaging was positioned adjacent to the midline and about 300 μm anterior of the PPC location. The location for M2 imaging was positioned one millimeter anterior of the RSC location.

Pre-processing of imaging data. We used custom code to correct for motion artifacts, as described in Ref. [113]. In brief, motion correction was implemented as a sum of shifts on three distinct temporal scales: sub-frame, full-frame, and minutes- to hour-long warping. After motion correction, ROIs were extracted using Suite2P.[114] Afterwards, somatic sources were identified with a custom two-layer convolutional network in MATLAB trained on manually annotated labels to classify ROIs as neural somata, processes, or other.[113] Only somatic sources were used. This yielded large populations from neurons from each of the four targeted regions in Mouse A and Mouse B (Table 5).

After identifying individual neurons, we computed average fluorescence in each ROI and converted this value into a normalized change in fluorescence (*ΔF/F*). We corrected the numerator of the *ΔF/F* calculation for neuropil by subtracting a scaled version of the neuropil signal estimated per neuron during source extraction:

$$F_{corr} = F - 0.7F_{neuropil} \tag{31}$$

We estimated the baseline fluorescence (F_{base}) of this trace as the 8th percentile of fluorescence within a 60-s window and subtracted this baseline to get the numerator:

$$\Delta F = F_{corr} - F_{base} \tag{32}$$

We divided this by the baseline (again 8th percentile of 60s window) of the raw fluorescence signal to get *ΔF/F*. We deconvolved the *ΔF/F* trace per neuron using the constrained AR-1 OASIS method44. We initialized the decay constants at two seconds and then optimized separately for each neuron. To fit the Model RNN, we temporally smoothed the sparse deconvolved spike estimates using a Gaussian kernel with four times the width of the sampling rate. We applied the same filter to the behavioral signals (pitch, roll, and yaw of the ball) to preserve the temporal relationship with the neural activity. For visualization of the neural population activity in the heatmaps of **Fig. 4 and S5**, we scaled each neuron by the mean of the total activity.

CURBD analysis to infer source currents from mouse data. For each mouse, we trained Model RNNs (Mouse A: 2787 units; Mouse B, Session 1: 999 units; Mouse B, Session 2: 1444 units) to reproduce the time-series Ca^{2+} data from the four regions. We used identical parameters for each Model RNN (Table 2).

We applied CURBD to infer the sixteen source currents comprising the multi-region population activity. We first assessed how much unique explanatory power each source current had in the total V1 population. We developed a partial coefficient of determination analysis to quantify this as follows. We subtracted each source current one-by-one from each V1 neuron and computed the sum-squared error of this difference and the recorded neural data. We then computed the sum-squared error of the full Model RNN fit compared to the recorded neural data. We defined the unique variance explained by the source current according to:

$$\text{VAF}_{unique} = 1 - \frac{\sum_t^T \left\{a(t) - [r(t) - I(t)]\right\}^2}{\sum_t^T [a(t) - r(t)]^2} \tag{33}$$

where *I(t)* denotes the partial source current that is being evaluated. Effectively, this computes the variance that cannot be explained by any of the three remaining source currents. Importantly, this metric can be computed at individual time points. We normalized each calculation by the sum of the four unique variances at each time point to give a proportion of unique variance explained by each source current. For cleaner presentation, we smoothed these normalized traces with a Gaussian kernel of width 500 ms (**Fig. 4h**).

We then reduced the dimensionality of all sixteen source currents using PCA, selecting a 5-dimensional manifold which sufficed to explain more than 80% of the total variance in all source currents. We trained Wiener cascade filters, a type of linear-nonlinear decoder,[115] to predict the running speed using the five-dimensional activity of each source current at each time step as well as the most recent 5 time steps of history. To perform cross validation, we randomly withheld 20% of time steps (the test set) and trained the decoders using the remaining 80% of the data. We quantified the performance of each decoder output on the left-out test set of time steps using VAF, as described above. We repeated this process for 100 iterations, randomly leaving out 20% of time steps for the test set on each iteration, and averaged across all iterations for the final decoder performance (**Fig. 4k,l**).

Multi-region electrophysiology recordings in monkeys

Behavioral task. All procedures were reviewed and approved by the Icahn School of Medicine Animal Care and Use Committee. For detailed descriptions of the experimental setup and protocol, see Ref. [9], where these data were previously reported. In brief, two rhesus macaque monkeys (*Macaca mulatta*; Monkey D: female, 5.6 kg; Monkey H: Male, 11.0 kg) were trained to sit in a custom primate chair with their head restrained and fixate on a computer monitor for four seconds, before performing a Pavlovian conditioning task for liquid rewards. They fixated on a neutral gray square for 800-1000ms. They were then presented with one of three visual conditioned stimuli for 500-600 ms on each trial corresponding to three different reward outcomes: no reward, water (0.5 mL), and juice (0.5 mL). An additional trial type occurred with equal frequency in which no conditioned stimuli was presented, and the gray square persisted throughout the trial. On all trials, a small (0.1 mL) water reward was given two seconds after the stimulus onset. Conditioned stimuli varied between monkeys and consisted of gray shapes, covering 1.10° of visual angle for Monkey D and 2.45° for Monkey H. We trained the Model RNNs described below using all four trial types to utilize as much training data as possible, though in this paper we only analyzed the CURBD output for the three conditioned stimuli.

Surgical procedures and neural recordings. After training, each monkey was implanted with a titanium head restraint device followed by a plastic recording chamber over the exposed cranium of the left frontal lobe. During the behavioral experiments, tungsten microelectrodes (FHC, Inc. or Alpha Omega, 0.5-1.5 M at 1 KHz) or 16-channel multi-contact linear arrays (Neuronexus Vector array) advanced by an 8-channel micromanipulator (NAN instruments, Nazareth, Israel) were attached to the recording chamber and inserted into the brain. The targeted brain regions were located using stereotaxic coordinates and verified by T1-weighted MRI imaging with the electrodes implanted. Recordings from subcallosal ACC were made on the medial surface of the brain ventral corpus callosum. Amygdala recordings were made between 22 and 18.5 mm anterior to the interaural plane. Rostromedial striatum recordings were made in the anterior medial segment corresponding to the zone where subcallosal and basal amygdala projections overlap. Spikes from putative single neurons were captured online using a Plexon Multichannel Acquisition Processor and later isolated with Plexon Offline Sorter. The small number of neurons recorded in each experimental session were then pooled into a pseudopopulation. First, the spike trains for each neuron on each trial were converted to an estimated firing rate. The firing rates were aligned on the stimulus presentation for each trial, then averaged across all trials of each stimulus type for that session in each monkey to give substantially large pseudopopulations (Table 6). As with the mouse dataset, neural activity was visualized with a heatmap after scaling each neuron's firing rate by its mean activity (**Fig. 5b and S7**).

CURBD analysis of monkey dataset. For each monkey, we trained Model RNNs (Monkey D: 343 units; Monkey H: 199 units) to match the pseudopopulation data for all four conditions. We used identical parameters for the Model RNNs for both monkeys (Table 2). In the previous simulations and the mouse dataset, the Model RNN learned a single dynamical system that reproduces the neural data based on one initial condition. However, here we have four different initial conditions corresponding to the four trial types, and we seek to learn a single dynamical system that reproduces all of them. To achieve this, we concatenated the time-series data from the four conditions and reset the state of the Model RNN to match the real neural data at the first time point of each new condition. We repeated each Model RNN fit an additional four times, yielding five runs in total, each starting from a different randomly initialized matrix $\boldsymbol{J_0}$ each time. We performed CURBD using each Model RNN to infer the nine source currents comprising the full multi-region activity. We then quantified the magnitude of current arriving to each source region from each target region by summing the absolute value of the source currents at each time step. We averaged this across the five runs in each monkey to assess the consistency of our solutions (**Fig. 5f,g**).

We performed a systematic subsampling analysis to assess whether applying CURBD to pseudopopulation data would be reliable even if different numbers and types of neurons were recorded experimentally. We randomly subsampled between 50% and 90% of the available neurons to create new, smaller pseudopopulations for each monkey. We used CCA (similar to the description in X above) to compute a similarity metric between the currents inferred by CURBD from each subsampled population and the currents originally inferred by CURBD using all of the available neurons for each monkey (**Fig. S8**).

Multi-region electrophysiology recordings in humans

Behavioral task. The institutional review boards of Cedars-Sinai Medical Center and the California Institute of Technology approved all protocols. Detailed descriptions of the experimental procedures are described in Ref. [10], where these data were previously reported. In brief, recordings were made from thirteen adult participants being evaluated for surgical treatment of drug-resistant epilepsy that provided informed consent and volunteered for this study. Of these thirteen participants, eight did not have a sufficiently large number of neurons to create a population for CURBD and were thus excluded. Our final analyses focused on five participants (P44, P51, P56, P57, and P58 from the original manuscript). The participants were seated in a chair facing a screen and reported decisions using either button presses or eye movements. They each performed eight forty-trial blocks that alternated between two tasks. In the categorization task, participants classified pseudorandomly presented images as belonging to one of four target categories (human faces, monkey faces, fruits, or cars) with a “yes” or “no” response. In the memory task, participants were shown an image and asked “Have you seen this image before, yes or no?” to which they responded “yes” or “no”. In the first block, all images were necessarily novel (40 unique images). In all subsequent blocks, the participants viewed 20 new images that were randomly intermixed with 20 familiar images. The 20 repeated images remained the same throughout the remainder of the session. We ignored trials where the participant provided an incorrect response (e.g. mistakenly identifying a novel image as familiar). This gave sixteen different conditions: four blocks of trials for each of two different tasks, each with correct “yes” and “no” responses. Participant task performance tended to improve throughout the session. Thus, we primarily focused our analysis on blocks 4, 6, and 8 (the final three memory blocks) when performance was highest.

Neural recordings. As participants performed this task, we recorded bilaterally from the amygdala, hippocampus, pre-supplementary motor area (preSMA), and dorsal anterior cingulate cortex (dACC) of each participant using microwires embedded in a hybrid electrode.[68] Electrode locations were confirmed using post-operative MRI or CT scans. We identified putative single neurons using a semi-automated spike sorting procedure. For the purposes of the CURBD analyses, we pooled recordings from each region from either hemisphere to ensure that we had sufficiently large populations for the dimensionality-reduction analysis. We estimated the instantaneous firing rate of each recorded neuron by convolving the spike train with a Gaussian kernel of width 150 ms. The relatively large width was necessary to accurately estimate firing rates for many low-firing neurons in the hippocampus and amygdala, and we opted to use a uniform width for all neurons. We created pseudopopulations by aligning each trial on the time of stimulus presentation and averaging across all trials for each task and correct response type (Table 7). In each trial, we kept two seconds before the stimulus presentation and three seconds after the stimulus presentation. As with the previous datasets, neural activity was visualized with a heatmap after scaling each neuron’s firing rate by its mean activity (**Fig. 5i and S9**).

CURBD analysis of human dataset. We trained Model RNNs based on the spiking pseudopopulation data from all sixteen conditions for each participant. As with the monkeys, we compensated for the discontinuities at trial boundaries by resetting the state of the Model RNN at the start of each condition. We used the same Model RNN parameters for all participants to ensure consistency (Table 2). We applied CURBD to infer the nine source currents comprising the multi-region interactions in the dataset. We reduced the dimensionality of each source current, as well as the full population activity of each region, using PCA. Since the stimulus was presented two seconds after the start of the trials, we defined the first two seconds to be the ‘resting state’ (**Fig. 5m**). We then computed the Mahalanobis distance of each source current at each time step. This gave a time-varying estimate of how much the population activity or source currents responded to the stimulus. For each participant, we averaged across all five training runs to obtain the most reliable estimate of the current dynamics. We then averaged across all participants to show the group effect (**Fig. 5n**).

REFERENCES

1. Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. (Barth, 1909).
2. Perich, M. G. & Rajan, K. Rethinking brain-wide interactions through multi-region ‘network of networks’ models. *Curr. Opin. Neurobiol.* **65**, 146–151 (2020).
3. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).

4. Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M. & Keller, P. J. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10**, 413–420 (2013).

5. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

6. Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent Network Models of Sequence Generation and Memory. *Neuron* **90**, 128–142 (2016).

7. Andalman, A. S. *et al.* Neuronal Dynamics Regulating Brain and Behavioral State Transitions. *Cell* **177**, 970–985.e20 (2019).

8. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986–999.e16 (2017).

9. Young, M. E. *et al.* Temporally-specific sequences of neural activity across interconnected corticolimbic structures during reward anticipation. *bioRxiv* (2020) doi:10.1101/2020.12.17.423162.

10. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, (2020).

11. Krubitzer, L. A. & Seelke, A. M. H. Cortical evolution in mammals: the bane and beauty of phenotypic variability. *Proc. Natl. Acad. Sci. U. S. A.* **109 Suppl 1**, 10647–10654 (2012).

12. Béna, G. & Goodman, D. F. M. Extreme sparsity gives rise to functional specialization. *arXiv [q-bio.NC]* (2021).

13. Tian, J. *et al.* Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons. *Neuron* **91**, 1374–1389 (2016).

14. Schröder, S. *et al.* Arousal Modulates Retinal Output. *Neuron* **107**, 487–495.e9 (2020).

15. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).

16. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255 (2019).

17. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).

18. Fisher, D., Olasagasti, I., Tank, D. W., Aksay, E. R. F. & Goldman, M. S. A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron* **79**, 987–1000 (2013).

19. Vogels, T. P., Rajan, K. & Abbott, L. F. Neural network dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).

20. Lovett-Barron, M. *et al.* Ancestral Circuits for the Coordinated Modulation of Brain State. *Cell* **171**, 1411–1423.e17 (2017).

21. Barak, O. Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* **46**, 1–6 (2017).

22. Yang, G. R. & Wang, X.-J. Artificial Neural Networks for Neuroscientists: A Primer. *Neuron* **107**, 1048–1070 (2020).

23. Sompolinsky, H., Crisanti, A. & Sommers, H. J. Chaos in random neural networks. *Phys. Rev. Lett.* **61**, 259–262 (1988).

24. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural*

Comput. **25**, 626–649 (2013).

25. Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
26. DePasquale, B., Cueva, C. J., Rajan, K., Escola, G. S. & Abbott, L. F. full-FORCE: A target-based method for training recurrent networks. *PLoS One* **13**, e0191527 (2018).
27. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
28. Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron* **98**, 1005–1019.e5 (2018).
29. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* **21**, 102–110 (2018).
30. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian Computation through Cortical Latent Dynamics. *Neuron* **103**, 934–947.e5 (2019).
31. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
32. Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A. & Scherberger, H. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 32124–32135 (2020).
33. Cohen, Z., DePasquale, B., Aoi, M. C. & Pillow, J. W. Recurrent dynamics of prefrontal cortex during context-dependent decision-making. *bioRxiv* (2020) doi:10.1101/2020.11.27.401539.
34. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
35. Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S. & Sussillo, D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Adv. Neural Inf. Process. Syst.* **32**, 15696–15705 (2019).
36. Miri, A. *et al.* Spatial gradients and multidimensional dynamics in a neural integrator circuit. *Nat. Neurosci.* **14**, 1150–1159 (2011).
37. Perich, M. G., Gallego, J. A. & Miller, L. E. A Neural Population Mechanism for Rapid Learning. *Neuron* **100**, 964–976.e7 (2018).
38. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* **102**, 249–259.e4 (2019).
39. Perich, M. G. *et al.* Motor cortical dynamics are shaped by multiple distinct subspaces during naturalistic behavior. *bioRxiv* (2020) doi:10.1101/2020.07.30.228767.
40. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without

movement. *Nat. Neurosci.* **17**, 440–448 (2014).

41. Abbott, L. F. & Dayan, P. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. (MIT Press, 2005).

42. Rajan, K., Abbott, L. F. & Sompolinsky, H. Stimulus-dependent suppression of chaos in recurrent neural networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **82**, 011903 (2010).

43. Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nat. Neurosci.* **23**, 260–270 (2020).

44. Veuthey, T. L., Derosier, K., Kondapavulur, S. & Ganguly, K. Single-trial cross-area neural population dynamics during long-term skill learning. *Nat. Commun.* **11**, 4057 (2020).

45. Das, A. & Fiete, I. R. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nat. Neurosci.* **23**, 1286–1296 (2020).

46. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).

47. Kim, T. H. *et al.* Long-Term Optical Access to an Estimated One Million Neurons in the Live Mouse Cortex. *Cell Rep.* **17**, 3385–3394 (2016).

48. Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L. & Tank, D. W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56**, 43–57 (2007).

49. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).

50. Gao, P. *et al.* A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* (2017) doi:10.1101/214262.

51. Gallego, J. A. *et al.* Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).

52. Abbott, L. F., Rajan, K. & Sompolinsky, H. Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks. in *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance* (ed. Ding M, G. D.) 65–82 (2011). doi:10.1093/acprof:oso/9780195393798.003.0004.

53. Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S. & Sussillo, D. Universality and individuality in neural dynamics across large populations of recurrent networks. *Adv. Neural Inf. Process. Syst.* **2019**, 15629–15641 (2019).

54. Vladimirov, N. *et al.* Light-sheet functional imaging in fictively behaving zebrafish. *Nat. Methods* **11**, 883–884 (2014).

55. Aoki, T. *et al.* Imaging of neural ensemble for the retrieval of a learned behavioral program. *Neuron* **78**, 881–894 (2013).

56. Beck, A. T., Steer, R. A., Kovacs, M. & Garrison, B. Hopelessness and eventual suicide: a 10-year prospective study of patients

hospitalized with suicidal ideation. *Am. J. Psychiatry* **142**, 559–563 (1985).

57. Maier, S. F. & Seligman, M. E. P. Learned helplessness at fifty: Insights from neuroscience. *Psychol. Rev.* **123**, 349–367 (2016).

58. Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife* **5**, (2016).

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

60. Vann, S. D., Aggleton, J. P. & Maguire, E. A. What does the retrosplenial cortex do? *Nat. Rev. Neurosci.* **10**, 792–802 (2009).

61. Ebbesen, C. L. & Brecht, M. Motor cortex — to act or not to act? *Nat. Rev. Neurosci.* **18**, 694–705 (2017).

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

63. Gremel, C. M. & Costa, R. M. Premotor cortex is critical for goal-directed actions. *Front. Comput. Neurosci.* **7**, 110 (2013).

64. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural Manifolds for the Control of Movement. *Neuron* **94**, 978–984 (2017).

65. Fortunato, C. *et al.* Nonlinear manifolds underlie neural population activity during behaviour. *bioRxiv* 2023.07.18.549575 (2023) doi:10.1101/2023.07.18.549575.

66. Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. The brain basis of emotion: a meta-analytic review. *Behav. Brain Sci.* **35**, 121–143 (2012).

67. Haber, S. N., Kim, K.-S., Mailly, P. & Calzavara, R. Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J. Neurosci.* **26**, 8368–8376 (2006).

68. Minxha, J., Mamelak, A. N. & Rutishauser, U. Surgical and Electrophysiological Techniques for Single-Neuron Recordings in Human Epilepsy Patients. *Neuromethods* 267–293 (2018) doi:10.1007/978-1-4939-7549-5_14.

69. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

70. Stevenson, I. H. *et al.* Functional connectivity and tuning curves in populations of simultaneously recorded neurons. *PLoS Comput. Biol.* **8**, e1002775 (2012).

71. Ahrens, M. B. *et al.* Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* **485**, 471–477 (2012).

72. Scott, B. B. *et al.* Imaging Cortical Dynamics in GCaMP Transgenic Rats with a Head-Mounted Widefield Macroscope. *Neuron* **100**, 1045–1058.e5 (2018).

73. Prevedel, R. *et al.* Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11**,

727–730 (2014).

74. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).

75. Javadzadeh, M. & Hofer, S. B. Dynamic causal communication channels between neocortical areas. *bioRxiv* (2021) doi:10.1101/2021.06.28.449892.

76. Seth, A. K., Barrett, A. B. & Barnett, L. Granger Causality Analysis in Neuroscience and Neuroimaging. *J. Neurosci.* **35**, 3293–3297 (2015).

77. Lawlor, P. N., Perich, M. G., Miller, L. E. & Kording, K. P. Linear-nonlinear-time-warp-poisson models of neural activity. *J. Comput. Neurosci.* **45**, 173–191 (2018).

78. Sirota, A., Csicsvari, J., Buhl, D. & Buzsáki, G. Communication between neocortex and hippocampus during sleep in rodents. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2065–2069 (2003).

79. Glaser, J. I., Whiteway, M., Cunningham, J. P., Paninski, L. & Linderman, S. W. Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations. *bioRxiv* (2020) doi:10.1101/2020.10.21.349282.

80. Linderman, S., Nichols, A., Blei, D., Zimmer, M. & Paninski, L. Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in C. elegans. *bioRxiv* (2019) doi:10.1101/621540.

81. Friston KJ, Harrison L, Penny W. Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).

82. Cao, X., Sandstede, B. & Luo, X. A Functional Data Method for Causal Dynamic Network Modeling of Task-Related fMRI. *Front. Neurosci.* **13**, 127 (2019).

83. Nguyen, J. P. *et al.* Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1074–81 (2016).

84. Schaffer, E. S. *et al.* Flygenvectors: The spatial and temporal structure of neural activity across the fly brain. *bioRxiv* 2021.09.25.461804 (2021) doi:10.1101/2021.09.25.461804.

85. Cisek, P. Resynthesizing behavior through phylogenetic refinement. *Atten. Percept. Psychophys.* **81**, 2265–2287 (2019).

86. Preuss, T. M. Do rats have prefrontal cortex? The rose-woolsey-akert program reconsidered. *J. Cogn. Neurosci.* **7**, 1–24 (1995).

87. Turner, M. H., Mann, K. & Clandinin, T. R. The connectome predicts resting-state functional connectivity across the Drosophila brain. *Curr. Biol.* **31**, 2386–2394.e3 (2021).

88. Falkner, A. L. *et al.* Hierarchical Representations of Aggression in a Hypothalamic-Midbrain Circuit. *Neuron* **106**, 637–648.e6 (2020).

89. Hultman, R. *et al.* Brain-wide Electrical Spatiotemporal Dynamics Encode Depression Vulnerability. *Cell* **173**, 166–180.e14 (2018).

90. Mu, Y. *et al.* Glia Accumulate Evidence that Actions Are Futile and Suppress Unsuccessful Behavior. *Cell* **178**, 27–43.e19 (2019).

91. Hosseini, M. *et al.* I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neurosci. Biobehav. Rev.* **119**, 456–467 (2020).

92. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).

93. Wilmes, K. A. & Clopath, C. Inhibitory microcircuits for top-down plasticity of sensory representations. *Nat. Commun.* **10**, 5055 (2019).

94. Bassett, D. S. & Bullmore, E. Small-World Brain Networks. *The Neuroscientist* **12**, 512–523 (2006).

95. Abbott, L. F. *et al.* The Mind of a Mouse. *Cell* **182**, 1372–1376 (2020).

96. Scheffer, L. K. *et al.* A connectome and analysis of the adult central brain. *Elife* **9**, (2020).

97. Pillow, J. W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).

98. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1099–1115.e8 (2018).

99. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).

100. Beyeler, A. *et al.* Divergent Routing of Positive and Negative Information from the Amygdala during Memory Retrieval. *Neuron* **90**, 348–361 (2016).

101. Jetti, S. K., Vendrell-Llopis, N. & Yaksi, E. Spontaneous activity governs olfactory representations in spatially organized habenular microcircuits. *Curr. Biol.* **24**, 434–439 (2014).

102. Bruno, A. M., Frost, W. N. & Humphries, M. D. Modular deconstruction reveals the dynamical and physical building blocks of a locomotion motor program. *Neuron* **86**, 304–318 (2015).

103. Gallego-Carracedo, C., Perich, M. G., Chowdhury, R. H., Miller, L. E. & Gallego, J. A. Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner. *bioRxiv* (2021) doi:10.1101/2021.05.31.446454.

104. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).

105. Pereira, T. D. *et al.* Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).

106. Dezfouli, A., Nock, R., Arabzadeh, E. & Dayan, P. Neural Network Poisson Models for Behavioural and Neural Spike Train Data. *bioRxiv* (2020) doi:10.1101/2020.07.13.201673.

107. Cao, R. & Yamins, D. Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *arXiv [q-bio.NC]* (2021).

108. Haykin, S. S. *Adaptive Filter Theory*. (Prentice Hall, 2002).

109. Safaie, M. *et al.* Preserved neural dynamics across animals performing similar behaviour. *Nature* (2023) doi:10.1038/s41586-023-06714-0.

110. Peng, Y., Ganesh, A., Wright, J., Xu, W. & Ma, Y. RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2233–2246 (2012).

111. Marshel, J. H., Garrett, M. E., Nauhaus, I. & Callaway, E. M. Functional specialization of seven mouse visual cortical areas. *Neuron* **72**, 1040–1054 (2011).

112. Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. *Neuron* **38**, 529–545 (2003).

113. Chettih, S. N. & Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in V1. *Nature* **567**, 334–340 (2019).

114. Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* (2017) doi:10.1101/061507.

115. Glaser, J. I. *et al.* Machine Learning for Neural Decoding. *eNeuro* **7**, (2020).

SUPPLEMENTAL FIGURES

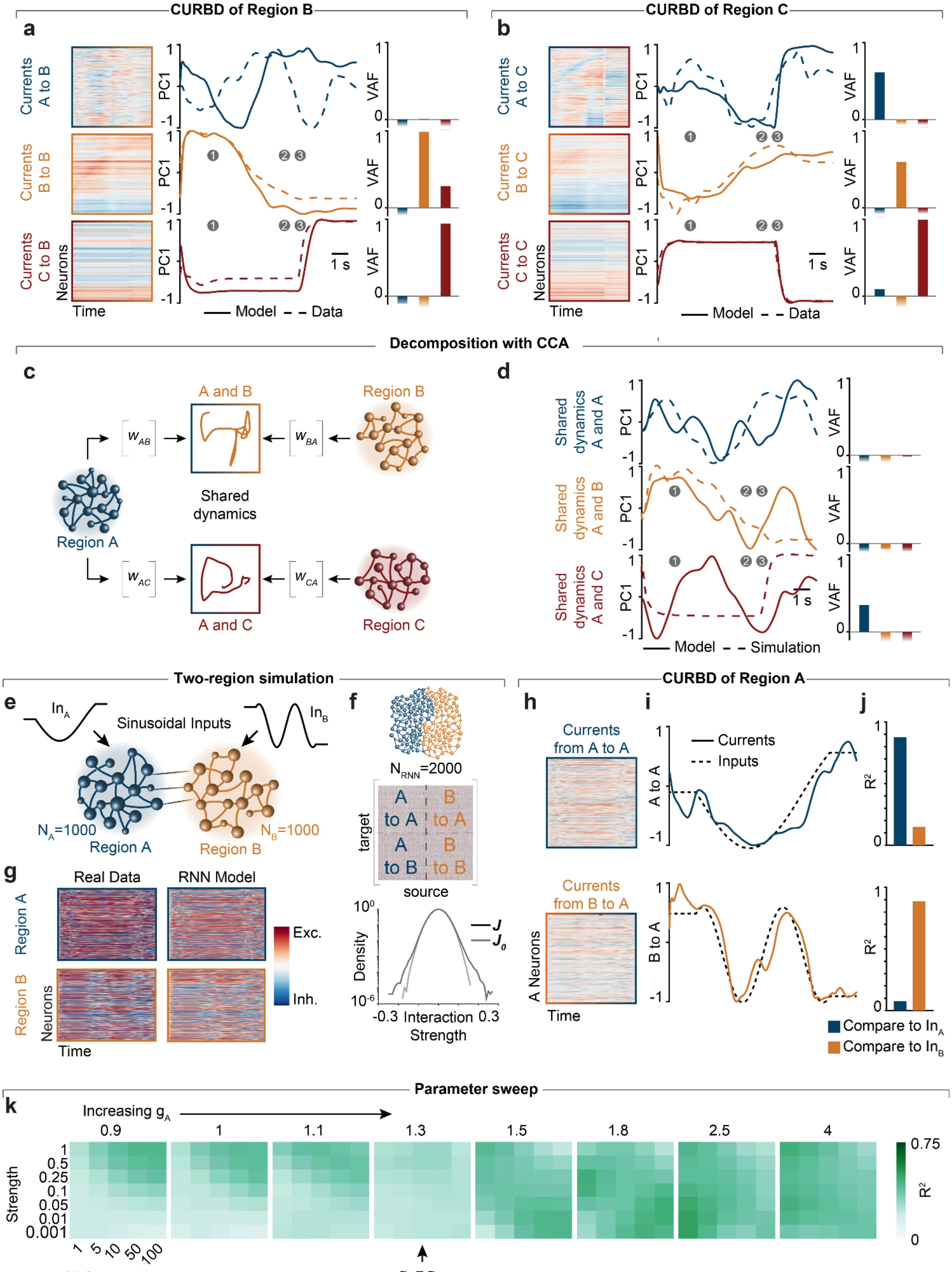
CURBD of Region B
a
Currents A to B
Currents B to B
Currents C to B
Neurons
Time
PC1
1
0
-1
VAF
1 s
Model
Data
CURBD of Region C
b
Currents A to C
Currents B to C
Currents C to C
Decomposition with CCA
c
A and B
Region B
w_{AB}
w_{BA}
Shared dynamics
Region A
w_{AC}
w_{CA}
A and C
Region C
d
Shared dynamics A and A
Shared dynamics A and B
Shared dynamics A and C
Model
Simulation
Two-region simulation
e
In_A
Sinusoidal Inputs
In_B
N_A=1000
Region A
Region B
N_B=1000
f
N_{RNN}=2000
A to A
B to A
A to B
B to B
target
source
10^0
10^{-6}
Density
J
J_0
-0.3
Interaction Strength
0.3
g
Real Data
RNN Model
Region A
Region B
Exc.
Inh.
Neurons
Time
CURBD of Region A
h
Currents from A to A
Currents from B to A
A Neurons
Time
i
Currents
Inputs
A to A
B to A
j
R^2
Compare to In_A
Compare to In_B
Parameter sweep
k
Increasing g_A
0.9
1
1.1
1.3
1.5
1.8
2.5
4
Strength
1
0.5
0.25
0.1
0.05
0.01
0.001
1
5
10
50
100
0.75
R^2
0

Fig. S1 | Supporting data for three-region ground truth simulations. a, Analysis of source currents within Region B. Data presented as in **Fig. 2g**. Currents from Region B and Region C are accurately reconstructed, though currents from Region A are missed, presumably due to the lack of strong external drive to Region A and the similar intrinsic dynamics between the three regions ($VAF_{AtoB}<0$; $VAF_{BtoB}=0.98$; $VAF_{CtoB}=0.94$). **b,** Analysis of source currents within Region C. Data presented as in **Fig. 2g**. All three source currents are accurately reconstructed ($VAF_{AtoC}=0.61$; $VAF_{BtoC}=0.60$; $VAF_{CtoC}=0.99$). **c,** CCA finds a single space capturing shared dynamics between each region, with a linear transformation (provided by the weight matrices *w*) relating each source and target region. However it does not provide a directional estimate of interactions. The shared dynamics plots show the aligned trajectories between pairs of regions projected onto the leading two aligned components. **d,** Comparison of ground truth current inputs and shared dynamics identified by CCA. Unlike CURBD, the shared dynamics identified by CCA do not accurately match the ground truth current dynamics ($VAF_{AandA}<0$; $VAF_{BandA}<0$; $VAF_{CandA}<0$). **e,** We simulated two interconnected RNNs representing distinct brain regions. Each was driven by a sinusoid of different frequencies. **f-g,** We fit a Model RNN directly to the time-series data of the two regions to perform CURBD. From the Model RNN we obtained a matrix describing the directed interactions within and between each of the two regions. **h,** We applied CURBD to obtain the currents driving each neuron in Region A from other Region A neurons (top) and from region B (bottom). **i,** We performed PCA to identify the dominant component of each source current. The currents from Region A resembled the low-frequency sinusoid driving Region A, while the currents from Region B matched the higher-frequency sinusoid driving Region B. **j,** We computed R^2 values comparing the first PC of each source current to the two sinusoidal inputs. **k,** Reconstruction accuracy of B to A currents for different simulation parameter values. We explored three key simulation parameters: i) the amount of chaos (g parameter; see Methods) from overdamped ($g<1$) to strongly chaotic ($g>1.5$); ii) the strength of the external inputs driving the system from very weak (0.001) to very strong (1); iii) the sparsity of inter-region connections from very sparse (1%) to full-rank (100%). Each heatmap shows the strength of inter-region connections against the percent of neurons receiving inter-region connections, and heatmaps going left to right show increasing g_A. For low values of g_A corresponding to damped dynamics, the inputs can only be reconstructed with strong connectivity. When both regions have similar dynamics ($g_A = 1.3$) the currents cannot be accurately demixed. The optimal regime occurs when g_A and g_B have different dynamics, with a tradeoff between sparsity and strength of the inter-region connectivity.

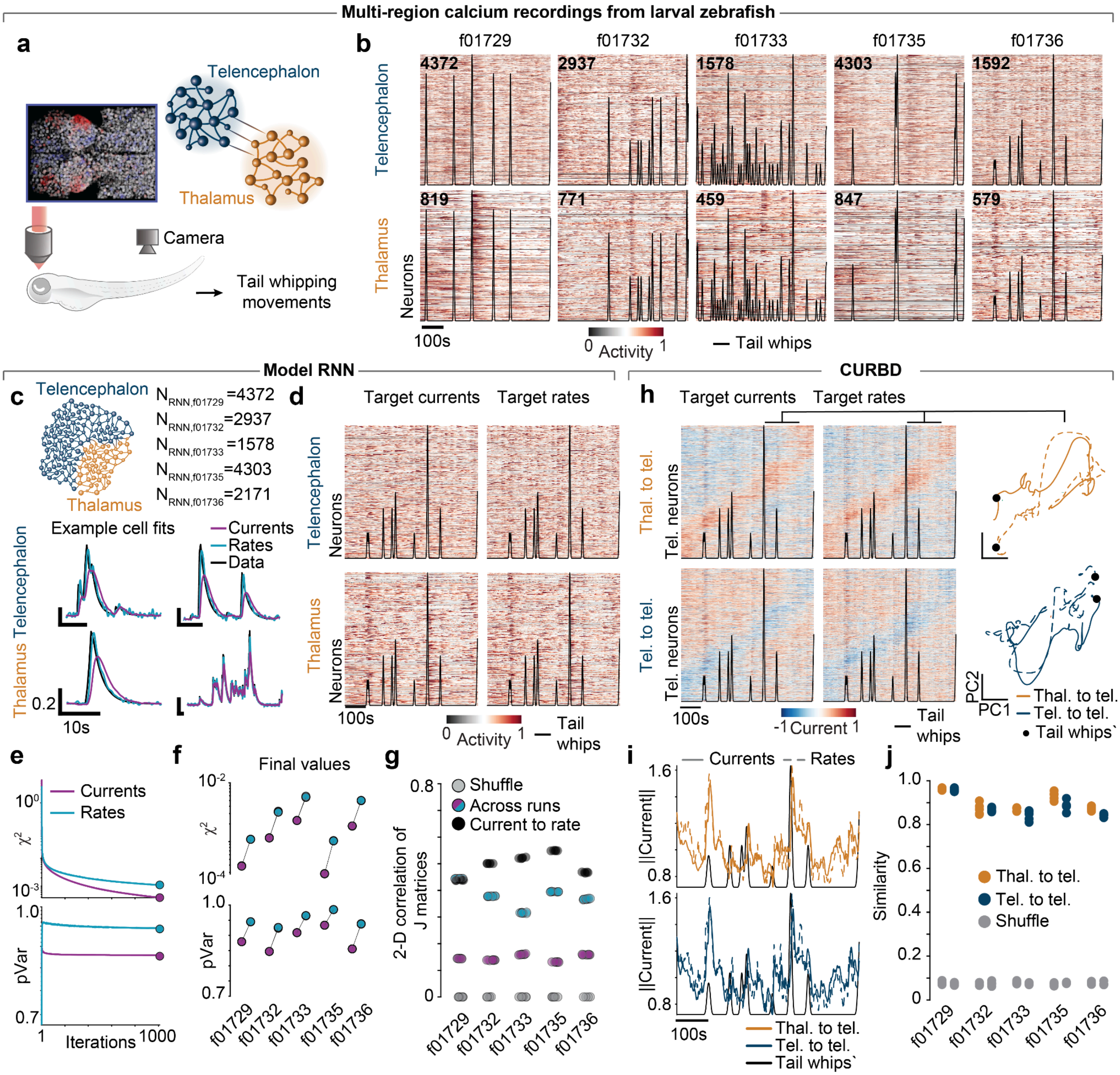

Fig. S2 | Comparing learning rules for Model RNNs using calcium recordings from larval zebrafish. a, Five larval zebrafish expressing GCaMP6s were head-fixed in order to image multi-region activity. The tails were kept free so that we could monitor tail whips during attempted swimming. The timing of tail movements was recorded with an overhead camera. **b,** Example activity of neural populations imaged from telencephalon and thalamus over time for the five fish. Redder colors indicate higher activity. Overlaid black lines indicate the presence of tail whipping movements. Inlaid numbers indicate the number of neurons recorded from each region. In each of the five fish, we observed phasic patterns of neural activity in each region, often coinciding temporally with bouts of tail movements. **c,** (top) We fit Model RNNs to recordings from telencephalon and thalamus for five fish. (bottom) Example fits to single neurons from telencephalon (top) and thalamus (bottom) using Model RNNs derived using current-based (red) and rate-based (blue) learning rules. **d,** Model RNN activity from f01736 for fits to the data sample shown in Panel b using current-based (left) and rate-based (right) learning rules. **e,** Example training performance curves from f01736 for the two learning rules. **f,** Final training performance for the two learning rules for Model RNNs fit to the five fish. The two learning rules gave similar fits, though rate-based learning was able to better capture high-frequency peaks in the neural data (see Panel C). This resulted in a larger variance explained (pVar) for the rate-based learning rule compared to the current-based learning rule. However, rate-based learning sometimes showed an increase in overall learning error when the target function values hovered around zero. Thus, there is an inherent tradeoff in the two learning rules. **g,** We tested the consistency of the quantities inferred from the Model RNN fits with the two learning rules. We computed the two-dimensional correlation of the J matrices, comparing the two learning rules on the same run (black), across runs with current-based (magenta) or rate-based (cyan) training, as well as a shuffled condition (gray). We found high similarity between the two rules ranging from 0.45 to 0.6. As a reference, we compared the similarity across runs for each learning rule and found that nearly all across-run values were less similar than the within-run values for the learning rules. Intriguingly, rate-based learning was more similar across runs, indicating that this approach may be capable of identifying more accurate and reliable solutions. **h,** We compared the dynamics of the inter-region currents inferred by CURBD using Model RNNs trained with both types of learning rules. (left) Heatmaps show inhibitory (blue) and excitatory (red) driving each telencephalon neuron over time. Neurons are sorted by the time of peak activation of excitatory thalamic inputs, and the same sort is used for the telencephalon inputs. (right) Projections into the leading two PCs to demonstrate the similarity in dynamics of currents derived from current (solid) and rate (dashed) training for a brief window of time indicated with the black bars. Black dots indicate the neural state at the time of each tail whipping motion. Both learning rules identified an intriguing balance of excitation and inhibition, with excitatory thalamic inputs coinciding with inhibitory local recurrent dynamics. **i,** Magnitude of current inputs estimated using current (solid) and rate (dashed) training. **j,** Similarity obtained using CCA (STAR Methods) between CURBD output using current and rate training for the five fish. The identified currents were highly similar (mean correlations > 0.8) for all five fish.

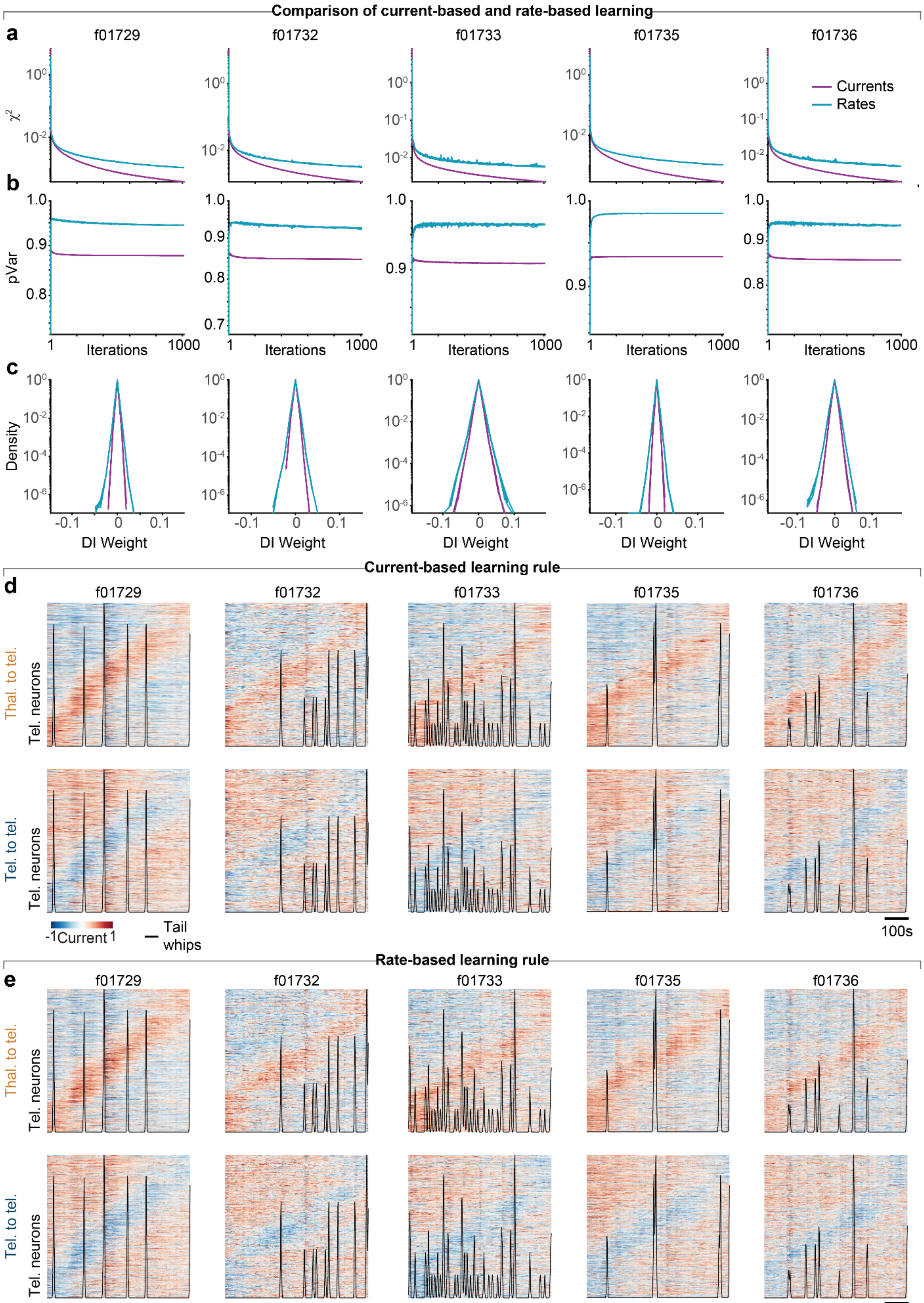
Comparison of current-based and rate-based learning
a
f01729
f01732
f01733
f01735
f01736
χ^2
10^0
10^{-1}
10^{-2}
Currents
Rates
b
pVar
1.0
0.9
0.8
0.7
1
Iterations
1000
c
Density
10^0
10^{-2}
10^{-4}
10^{-6}
-0.1
0
0.1
DI Weight
Current-based learning rule
d
f01729
f01732
f01733
f01735
f01736
Thal. to tel.
Tel. neurons
Tel. to tel.
Tel. neurons
-1
Current
1
Tail whips
100s
Rate-based learning rule
e
f01729
f01732
f01733
f01735
f01736
Thal. to tel.
Tel. neurons
Tel. to tel.
Tel. neurons

Fig. S3 | Supporting data for comparing current-based and rate-based Model RNNs in the larval zebrafish dataset. a, Model error over 1000 training iterations for the Model RNN for the five fish for current-based learning (magenta) and rate-based learning (cyan). **b,** Proportion of variance explained by the Model RNN for 1000 training iterations for the two learning rules. **c,** Final distributions of directed interaction weights for the trained **J** matrices using the two learning rules. **d,** CURBD on the inputs to telencephalon neurons from thalamus (top) and recurrent connections from telencephalon (bottom) using the current-based learning rule. Heatmaps show inhibitory (blue) and excitatory (red) driving each telencephalon neuron over time. Neurons are sorted by the time of peak activation of excitatory thalamic inputs, and the same sort is used for the telencephalon inputs. Black overlaid lines show tail whipping movements. **e,** Data presented as in Panel a for CURBD using the rate-based learning rule.

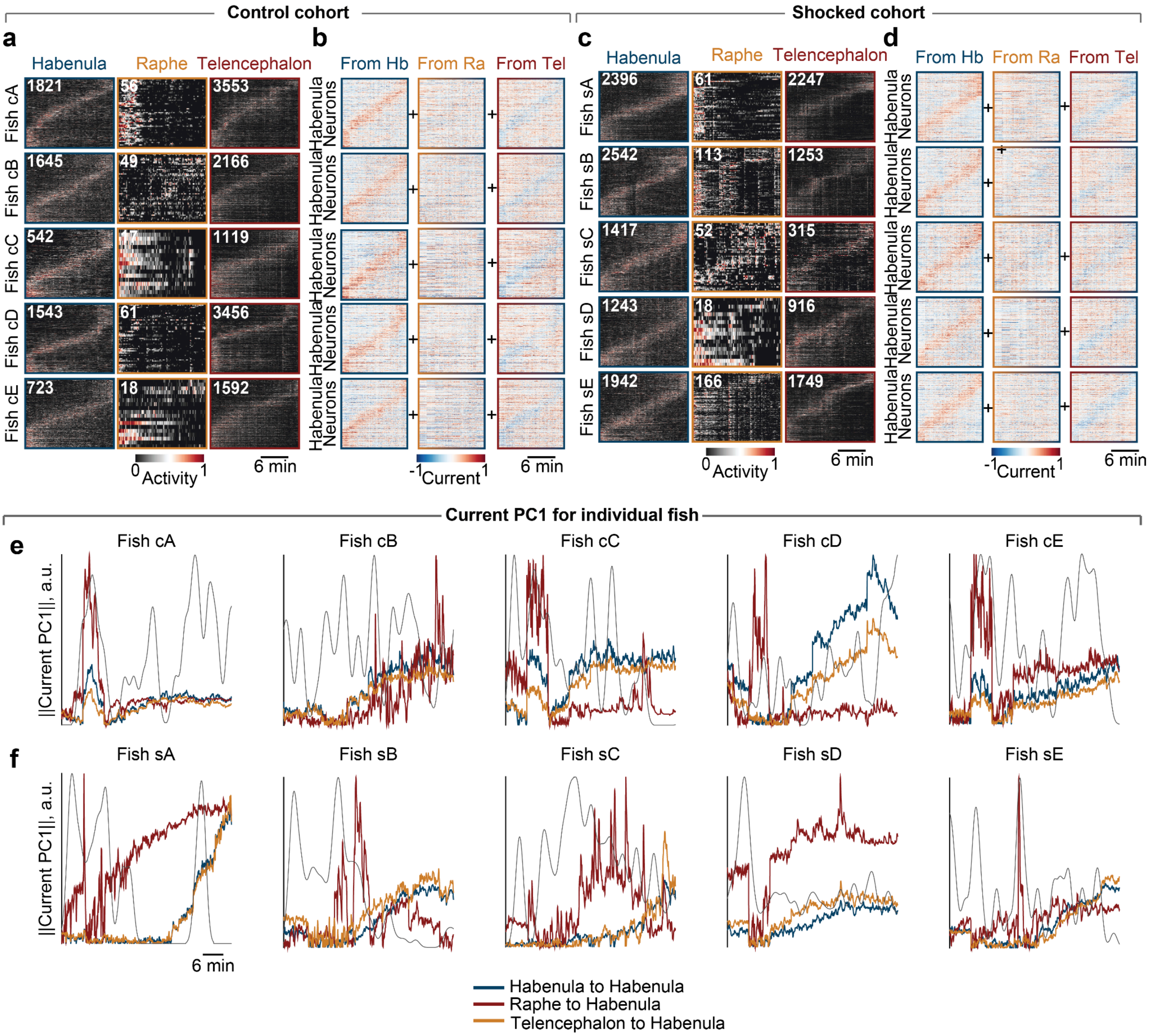

Fig. S4 | Supporting data for the three-region larval zebrafish dataset. a, Heatmaps of all recorded cells for the 3 regions in each individual fish of the control cohort. Data presented as in **Fig. .b,** Heatmaps of current inputs to habenular neurons for each fish of the control cohort. Data presented as in **Fig. 3h**. **c,** Same as Panel A for the shocked cohort. Panel duplicated from **Fig. 3b**. **d,** Same as Panel B for the individual fish of the shocked cohort. Top row (Fish sA) is reproduced from **Fig. 3h**. **e-f,** The leading Principal Component across Habenular neurons capturing the dominant pattern of inputs to Habenula from each region. Plots show each individual fish from the control (Panel e) and shocked (Panel f) cohorts.

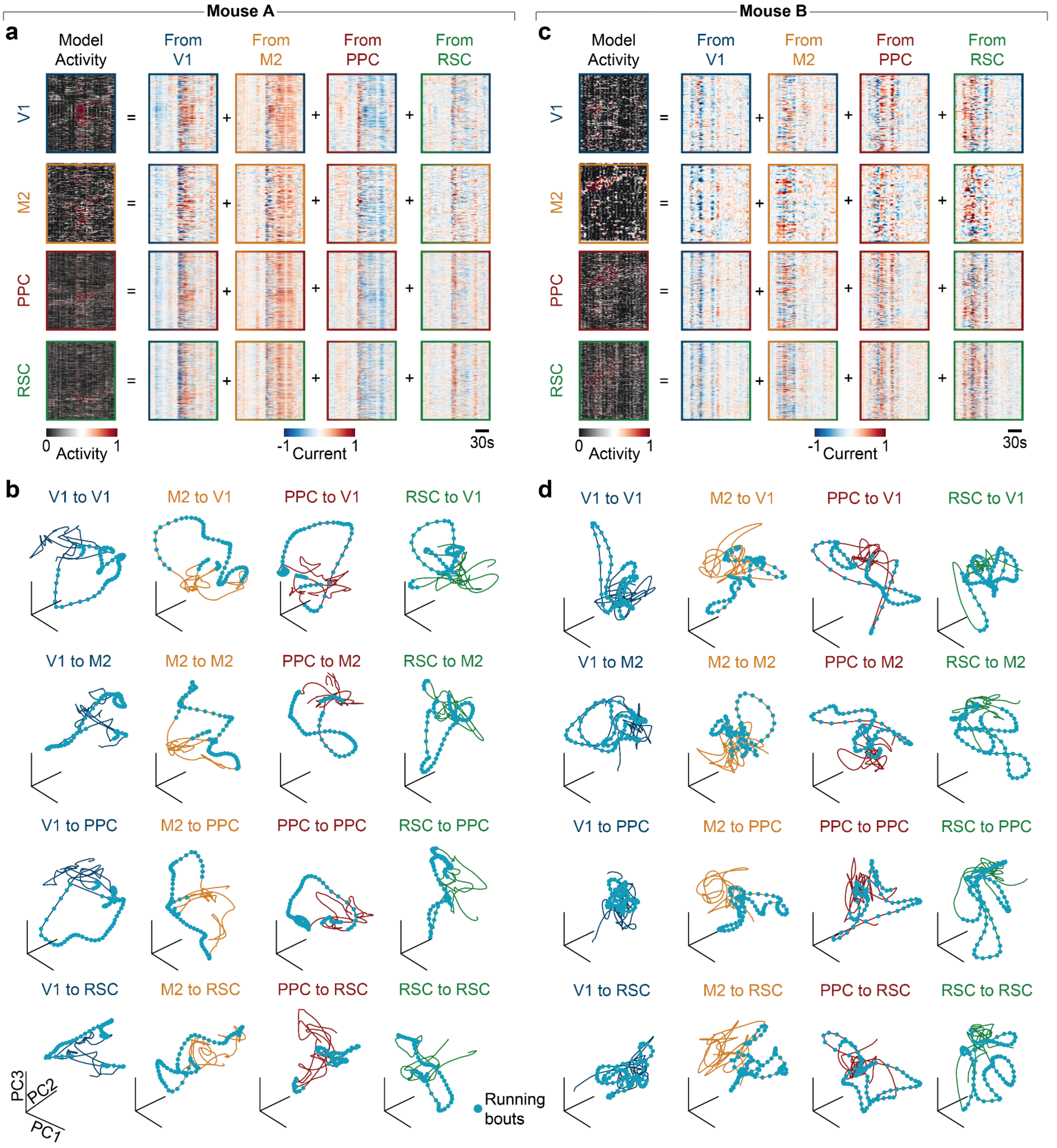

Fig. S5 | Supporting data for the multi-region mouse dataset. a, Model RNN activity and source current activity for Mouse A. Figures reproduced from **Fig. 4g**. **b,** Current trajectories in the first three PCs for all sixteen source currents from Mouse A. The V1 source currents (top row) are reproduced from **Fig. 4h**. **c,** Data presented as in Panel A for Mouse B. **d,** Data presented as in Panel B for Mouse B.

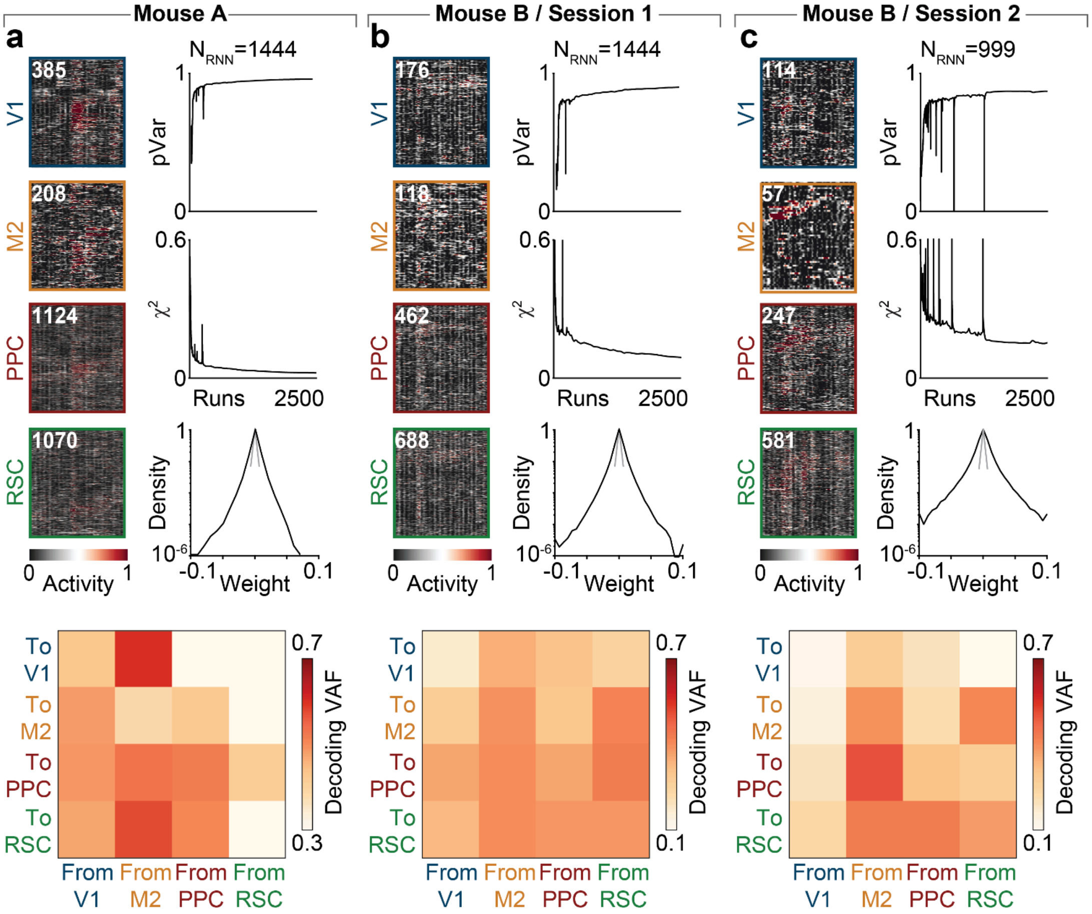

Fig. S6 | All recording sessions for the mouse dataset. a, (Top) Model RNN output (left) and training performance (right) for the session from Mouse A. (Bottom) Decoding performance for the sixteen source currents. All data are reproduced from **Fig. 4**. **b,** (Top) Model RNN output (left) and training performance (right) for Session 1 from Mouse B. (Bottom) Decoding performance for the sixteen source currents. Portions are reproduced from **Fig. 4**. **c,** (Top) Model RNN output (left) and training performance (right) for Session 2 from Mouse B. (Bottom) Decoding performance for the sixteen source currents.

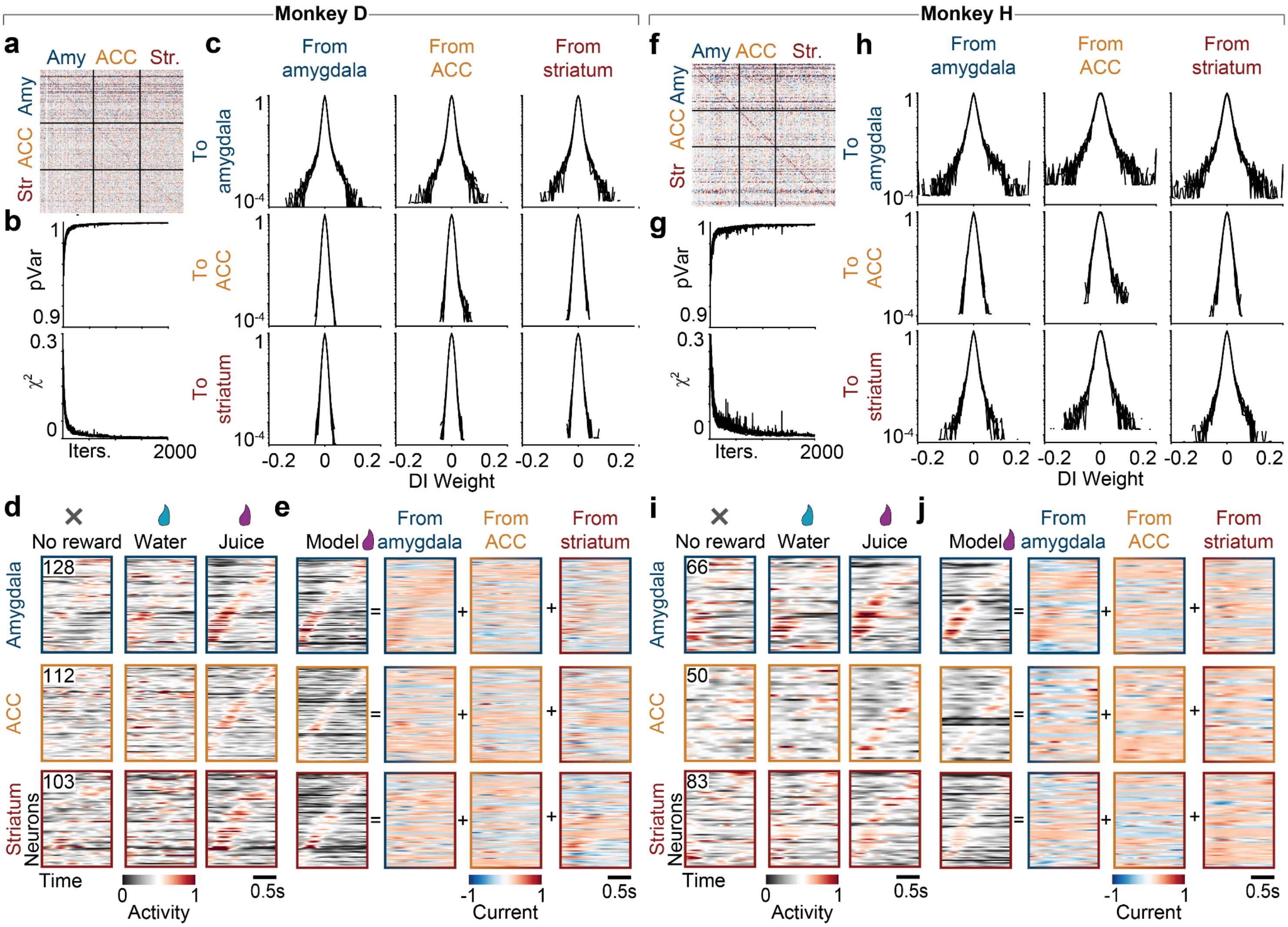

Fig. S7 | Supporting data for the multi-region macaque electrophysiology dataset. a, Connectivity matrix for an example Model RNN fit to data from Monkey D. Neurons are ordered by region, starting with amygdala (Amy, blue), subcallosal anterior cingulate cortex (ACC, yellow), and rostromedial striatum (Str, red). **b,** (Top) Proportion of variance explained (pVar) in the neural population as a function of training runs. Training results for five different random initializations are plotted to highlight consistency. (Bottom) Model error (χ^2) for the five initializations shown above. **c,** Distribution of weights in each submatrix used for CURBD. Each column corresponds to a source region, and each row to a target region. All five initializations are plotted to illustrate consistency. **d,** Trial-averaged firing rates for the amygdala (top), subcallosal ACC (middle), and striatum (bottom) comprising the pseudopopulation dataset for Monkey D. Left plot shows data from the unconditioned stimulus (left), water stimulus (middle), and juice stimulus (right). All trials are aligned on presentation of the stimulus. Neurons in each region are sorted according to their time of peak activity in the Juice condition. **e,** CURBD decomposition of activity in each region for the Juice trials. Left plots show the full Model RNN activity. The remaining plots show the inferred source currents to each target region (rows) from all source regions (columns). **f-j,** Data for Monkey H presented as in Panels A-E.

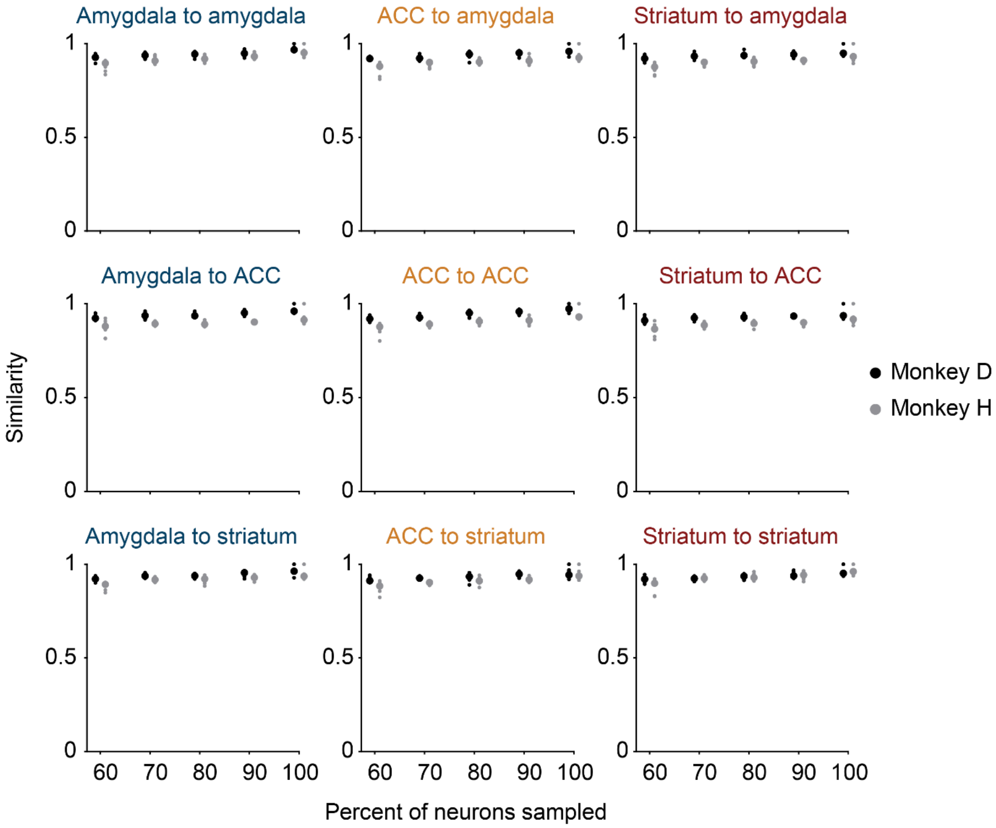

Fig. S8 | Consistent identification of current dynamics with random subsamples of recorded neurons. Mean canonical correlation in a 20-dimensional space identified by PCA for each source current in Monkey D (black) and Monkey H (gray). Small dots indicate the results of 10 random subsamples of the total neural population at each percentage level. Large circles indicate the median across iterations.

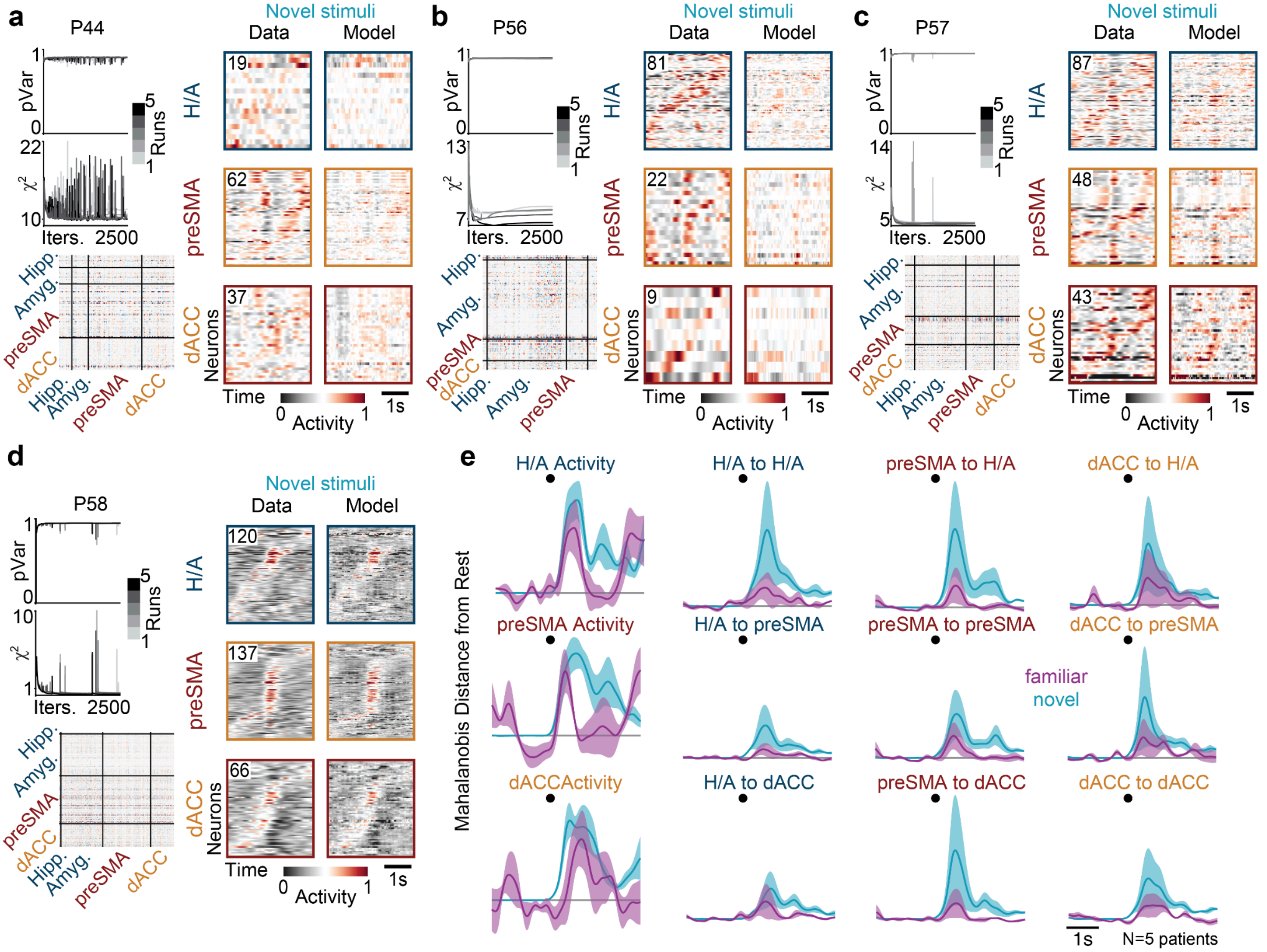

Fig. S9 | Supporting data for the multi-region human electrophysiology dataset. a, Model RNN summary for P44. (Top left) Model RNN training performance (pVar and χ^2) for five runs starting from different random initializations of the J matrix. (Bottom left) Example J matrix for one run. (Right) Neural activity from recorded neurons (Data) and the Model RNN units. **b,** Data presented as in Panel a for P56. **c,** Data presented as in Panel a for P57. **d,** Data presented as in Panel a for P58. **e,** Mahalanobis distance from rest for all sixteen source currents on the familiar stimuli trials (magenta) and novel stimuli trials (cyan). Lines show mean and standard error across all five participants. Black dot indicates time of stimulus presentation. Top row is reproduced from **Fig. 5n**.

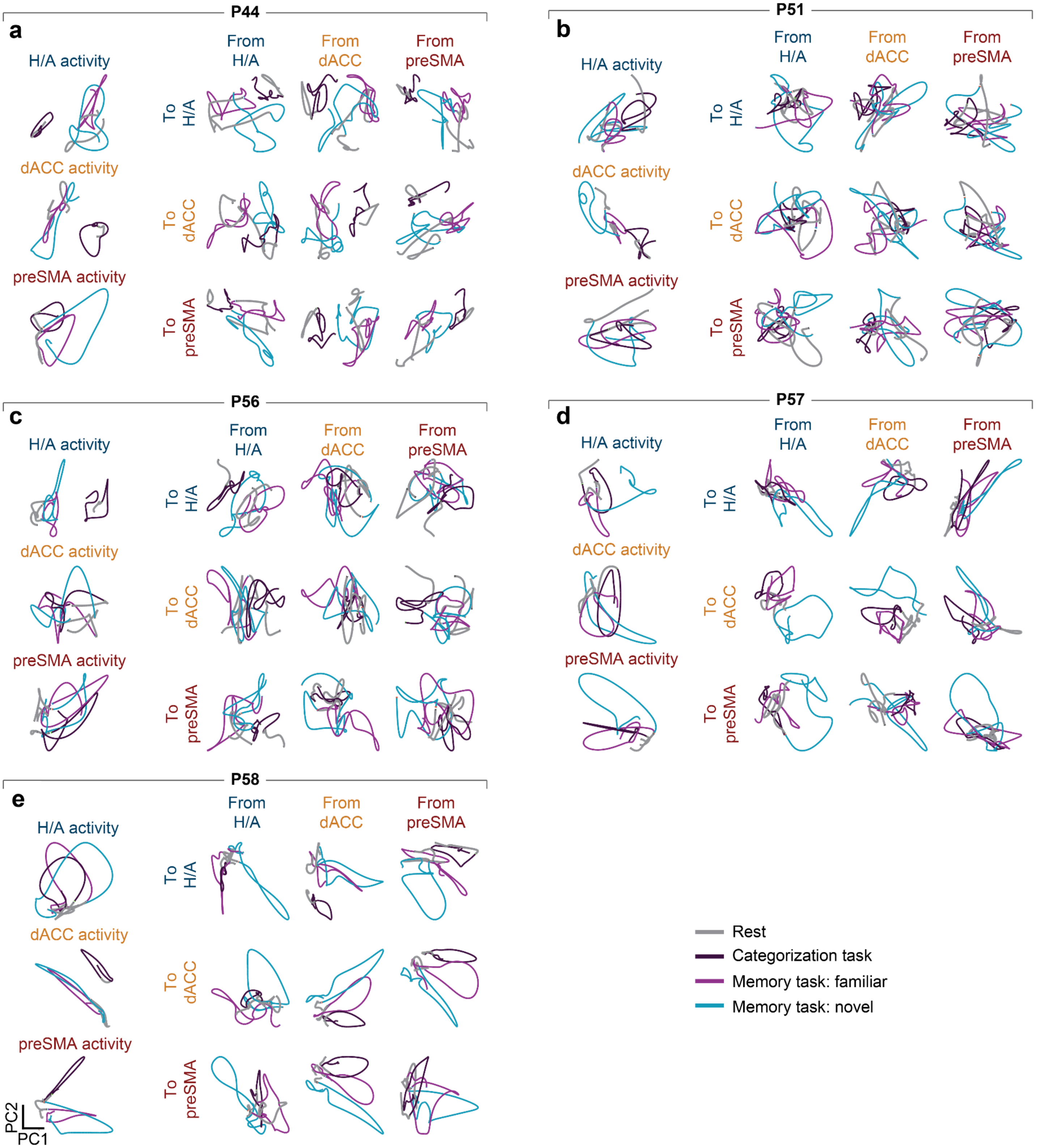

Fig. S10 | Comparison of current dynamics across tasks for the multi-region human electrophysiology dataset. a, Neural and current dynamics for both tasks in P44. Each subplot shows the first two PCs of the full population activity of the three regions as well as the nine source currents during the categorization task (purple) and novel (cyan) and familiar (magenta) stimuli during the memory task. Gray shows activity at rest before the stimuli. **b,** Neural and current dynamics for both tasks in P51. **c,** Neural and current dynamics for both tasks in P56. **d,** Neural and current dynamics for both tasks in P57. **e,** Neural and current dynamics for both tasks in P58.

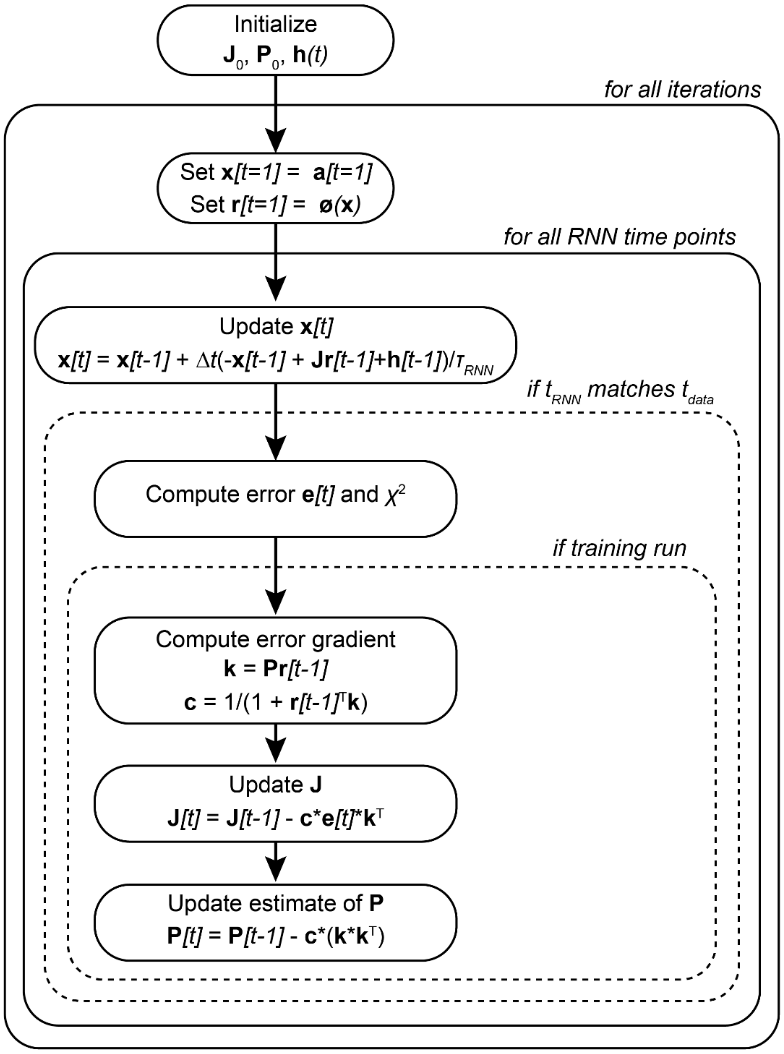
Initialize
$\mathbf{J}_0$, $\mathbf{P}_0$, $\mathbf{h}(t)$
for all iterations
Set $\mathbf{x}[t=1] = \mathbf{a}[t=1]$
Set $\mathbf{r}[t=1] = \mathbf{ø}(\mathbf{x})$
for all RNN time points
Update $\mathbf{x}[t]$
$\mathbf{x}[t] = \mathbf{x}[t\text{-}1] + \Delta t(-\mathbf{x}[t\text{-}1] + \mathbf{J}\mathbf{r}[t\text{-}1]+\mathbf{h}[t\text{-}1])/\tau_{RNN}$
if t_{RNN} matches t_{data}
Compute error $\mathbf{e}[t]$ and χ^2
if training run
Compute error gradient
$\mathbf{k} = \mathbf{P}\mathbf{r}[t\text{-}1]$
$\mathbf{c} = 1/(1 + \mathbf{r}[t\text{-}1]^{T}\mathbf{k})$
Update $\mathbf{J}$
$\mathbf{J}[t] = \mathbf{J}[t\text{-}1] - \mathbf{c}{*}\mathbf{e}[t]{*}\mathbf{k}^{T}$
Update estimate of $\mathbf{P}$
$\mathbf{P}[t] = \mathbf{P}[t\text{-}1] - \mathbf{c}{*}(\mathbf{k}{*}\mathbf{k}^{T})$

Fig. S11 | Flowchart of algorithmic implementation of learning rules. This flowchart indicates the algorithm to iteratively estimate the **P** matrix used during learning.

SUPPLEMENTAL TABLES

Table 1 | Two-region generator model parameters

Parameter	Description	Value(s)
g_A	Region A chaos	[0.9, 1.0, 1.1, 1.3, 1.5, 1.8, 2.5]
g_B	Region B chaos	1.5
τ	Time constant of model units	0.1
T	Simulation time	10

Parameter	Description	Value(s)
w_{rgn}	Inter-region weight	[0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 1.0]
p_{rgn}	Inter-region connection proportion	[0.01, 0.05, 0.1, 0.25, 0.5, 1.0]
w_{in}	External input weight	1.0
Δt	Simulation time step size	0.01

Table 2 | Model RNN training parameters for all datasets

Parameter	Description	2-region simulation	3-region simulation	Zebrafish dataset	Mouse dataset	Monkey dataset	Human dataset
g	Chaotic spontaneous activity	1.5	1.5	1.3	1.5	1.5	1.5
τ	Time constant for model units	0.1	0.1	2.5	0.3	0.001	0.0075
P_0	Learning rate	1.0	1.0	1.0	1.0	1.0	1.0
τ_{WN}	Time constant of filtered white noise	0.1	0.1	0.1	0.1	0.1	0.1
w_{WN}	White noise input weight	0.01	0.01	0.01	0.001	0.0001	0.001
$n_{iterations}$	Number of training iterations	100	500	1000	2500	1500	2500
Δt	Data time step size (s)	0.01	0.01	0.92 - 0.97	0.1866	0.01	0.01
Δt_{RNN}	Model RNN step size (s)	0.001	0.001	0.23 - 0.24	0.0467	0.0005	0.001

Table 3 | Three-region generator model parameters

Parameter	Description	Value
g_A	Region A chaos	1.8
g_B	Region B chaos	1.5
g_C	Region C chaos	1.5
τ_{true}	True decay constant	0.1
T	Simulation time	12

Parameter	Description	Value
w_{rgn}	Inter-region connection weight	0.01
p_{rgn}	Fraction of inter-region connections	0.01
w_{in}	External input weight	1.0
σ	Width of sequential and fixed point-bumps (number of units)	200
Δt	Simulation time step	0.01

Table 4 | Simultaneous neuron yield for the selected from the larval zebrafish dataset

Brain Region	**Fish cA**	**Fish cB**	**Fish cC**	**Fish cD**	**Fish cE**	**Fish sA**	**Fish sB**	**Fish sC**	**Fish sD**	**Fish sE**
Telenc.	3553	2166	1119	3456	1592	2247	1253	315	916	1749
Thalamus	819	771	459	847	579	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
Habenula	1821	1645	542	1543	723	2396	2542	1417	1243	1942
Raphe	56	49	17	61	18	61	113	52	18	166
Total	6249	4631	2137	5907	2912	4704	3908	1784	2177	3857

Table 5 | Simultaneous neuron yield for the mouse dataset

Brain Region	**Mouse A**	**Mouse B**	
		Session 1	*Session 2*
V1	385	114	176
M2	208	57	118
PPC	1124	247	462
RSC	1070	581	688
Total	2787	999	1444

Table 6 | Pseudopopulation sizes for the monkey dataset

Brain Region	**Monkey D**	**Monkey H**
Amygdala	128	66
Subcallosal ACC	112	50
Striatum	103	83
Total	343	199

Table 7 | Pseudopopulation sizes for the human dataset, reported as: Left Hemisphere / Right Hemisphere (Total)

Brain Region	**P44** (two sessions)	**P51** (five sessions)	**P56** (three sessions)	**P57** (three sessions)	**P58** (three sessions)
Hippocampus	0 / 15 (15)	25 / 38 (63)	5 / 0 (5)	17 / 0 (17)	5 / 0 (5)
Amygdala	3 / 1 (4)	2 / 22 (24)	33 / 43 (76)	36 / 34 (70)	54 / 61 (115)
preSMA	62 / 0 (62)	16 / 7 (23)	8 / 14 (22)	48 / 0 (48)	65 / 72 (137)

dACC	37 / 0 (37)	72 / 17 (89)	0 / 9 (9)	43 / 0 (43)	7 / 59 (66)
Total	102 / 16 (118)	115 / 84 (199)	46 / 66 (112)	144 / 34 (178)	131 / 192 (323)