Accurate Identification of Communication Between Multiple Interacting Neural Populations

Belle Liu[1] **Jacob Sacks**[2] **Matthew D. Golub**[2]

Abstract

Neural recording technologies now enable simultaneous recording of population activity across multiple brain regions, motivating the development of data-driven models of communication between recorded brain regions. Existing models can struggle to disentangle communication from the effects of unrecorded regions and local neural population dynamics. Here, we introduce Multi-Region Latent Factor Analysis via Dynamical Systems (MR-LFADS), a sequential variational autoencoder composed of region-specific recurrent networks. MR-LFADS features structured information bottlenecks, data-constrained communication, and unsupervised inference of unobserved inputs—features that specifically support disentangling of inter-regional communication, inputs from unobserved regions, and local population dynamics. MR-LFADS outperforms existing approaches at identifying communication across dozens of simulations of task-trained multi-region networks. Applied to large-scale electrophysiology, MR-LFADS predicts brain-wide effects of circuit perturbations that were not seen during model fitting. These validations on synthetic and real neural data suggest that MR-LFADS could serve as a powerful tool for uncovering the principles of brain-wide information processing.

1. Introduction

Large-scale neural recording technologies, such as high-density electrophysiology (Steinmetz et al., 2019; Siegle et al., 2021; IBL et al., 2023; Chen et al., 2024; Bennett et al., 2024) and calcium imaging (Sofroniew et al., 2016; Song et al., 2017; Allen et al., 2017), have enabled the study of simultaneously recorded brain-wide activity, revealing that many sensory, motor and cognitive processes engage spatially distributed networks of brain regions (Makino et al., 2017; Gilad et al., 2018; Stringer et al., 2019; Musall et al., 2019; Allen et al., 2019; Jia et al., 2022). Consequently, there has been growing interest in the design of data-driven *communication models* that seek to infer the pathways and content of communication between the recorded regions.

Accurately identifying such communication is challenging for at least *four* reasons (Biswas et al., 2020; Kang & Druckmann, 2020; Keeley et al., 2020; Perich & Rajan, 2020; Semedo et al., 2020; Kass et al., 2023). *First*, communication signals are not directly observed in multi-region recordings. While some of the recorded neurons might project to other recorded regions, the identity and targets of these projection neurons are typically unknown. *Second*, models may need to account for inputs to the recorded brain regions from other regions that were not recorded during the experiment. *Third*, models should faithfully reconstruct activity in each recorded region by accounting for inferred inter-regional communication, inputs from unrecorded areas, and local dynamics—while also capturing complex features such as structured trial-to-trial variability (Goris et al., 2014) and nonlinear, nonstationary, state-dependent population dynamics (Shenoy et al., 2013; Vyas et al., 2020; Duncker & Sahani, 2021; Durstewitz et al., 2023). *Fourth*, accurate reconstruction of the recorded data does not necessarily imply an accurate account of the underlying communication. Many different models can often sufficiently reconstruct the recorded data, leading to ambiguities in determining which, if any, of those models should be trusted as tools for scientific inquiry.

In this work, we introduce Multi-Region Latent Factor Analysis via Dynamical Systems (**MR-LFADS**), a multi-region communication model that directly addresses all of the challenges outlined above. MR-LFADS is a probabilistic model that represents each recorded region with a distinct set of stacked recurrent neural networks (**RNNs**) that capture the region's potentially nonlinear and nonstationary population dynamics. MR-LFADS represents communication between observed regions and inputs from unobserved regions as

[1] Graduate Program in Neuroscience, University of Washington [2] Paul G. Allen School of Computer Science & Engineering, University of Washington. Correspondence to: Matthew Golub <mgolub@cs.washington.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

disentangled sets of latent variables. Structured information bottlenecks encourage the model to infer inputs from unobserved regions only when their effects cannot be explained by communication from the recorded regions. MR-LFADS infers single-trial initial conditions and time-varying inputs that together account for trial-to-trial variability in the recorded activity. This automatic inference of inputs eliminates the need to manually specify input signals, thereby avoiding strong, hard-to-validate assumptions about how external signals influence each region. Finally, MR-LFADS constrains communication to originate from reconstructed neural activity in the upstream regions, rather than from more-flexible latent representations (Glaser et al., 2020; Karniol-Tambour et al., 2024)—a design choice that, as we show, enables more accurate inference of communication without sacrificing the quality of data reconstruction.

To evaluate MR-LFADS, we developed 37 synthetic multi-region datasets that pose the real-world challenges outlined above across a range of neuroscience-relevant scenarios for communication modeling. Here, MR-LFADS consistently outperforms existing models in recovering the pathways and content of communication. Through selective ablations of MR-LFADS design features, we show that these features indeed improve the identification of communication in these settings. Finally, we apply MR-LFADS to multi-region electrophysiology recordings in mice performing a decision-making task (Chen et al., 2024). In a subset of trials that were held out during model fitting, photoinhibition was applied to the anterior lateral motor cortex (**ALM**). We show that MR-LFADS predicts the brain-wide effects of these circuit perturbations, suggesting that MR-LFADS inferred an accurate account of inter-regional communication. Moreover, MR-LFADS infers consistent communication across multiple models trained from different random initializations, demonstrating its reliability and robustness in real-data settings.

2. Related Work

Existing communication models can be broadly categorized as either *static* or *dynamic* methods. *Static methods*, such as reduced-rank regression (**RRR**) (Semedo et al., 2019; Steinmetz et al., 2019; MacDowell et al., 2025), predict activity in a target region from simultaneous activity in one or more source regions (Kaufman et al., 2014; Ruff & Cohen, 2019; Veuthey et al., 2020) and then interpret the predictive source activity as communication. While these methods are straightforward to fit and interpret, they treat each time point independently and thus do not explicitly model temporal structure in neural recordings. *Dynamic methods* explicitly model the temporal structure in multi-region data. These approaches address temporal structure using, for example, switching linear dynamical systems (**SLDS**) models (Linderman et al., 2016; Glaser et al., 2020), RNNs (Perich et al., 2020; Karniol-Tambour et al., 2024), or Gaussian processes (Yu et al., 2008; Gokcen et al., 2022; 2024).

All of the aforementioned approaches meet some of the challenges outlined in Section 1, but, in our view, none of the approaches meet all of the challenges. Critically, none of these existing communication models support inferring inputs from unobserved brain regions. This functionality is critical because inputs from unobserved regions might modulate the target region's population dynamics and how that region communicates with other regions. Instead, some approaches compensate by manually specifying each region's inputs (e.g., as stimulus or other task-related signals) or by subtracting off condition averages from single-trial neural activity. However, these strategies impose strong assumptions about the content of input signals and the regions they target. As we will show, misspecifying such inputs risks confounding inferred population dynamics and communication.

Pandarinath et al. (2018) introduced **LFADS**, a technique for inferring unobserved time-varying inputs when modeling single-trial neural population dynamics within a single recorded brain region. LFADS is a sequential variational autoencoder **sVAE** that jointly identifies a nonlinear dynamical system, implemented as an RNN, along with the single-trial initial conditions and time-varying unobserved inputs needed to drive the system to reconstruct single-trial neural population recordings. We henceforth use **SR-LFADS** to refer to a single-region LFADS model.

3. Multi-Region LFADS (MR-LFADS)

MR-LFADS is composed of a set of SR-LFADS models (Fig. 1a) that interact through constrained communication channels (Fig. 1b). At a high level, MR-LFADS is a coupled set of driven nonlinear dynamical systems that are jointly trained to reconstruct all single trials of a multi-region dataset. Each recorded brain region i is represented by its own dynamical system, which attempts to reconstruct the region-i recorded neural activity x_t^i at each time $t = 1, \ldots, T$. Each region i dynamical system evolves from a single-trial initial state g_0^i and is driven by (1) single-trial time-varying communication messages $m_t^{j \to i}$ from other recorded brain regions j and (2) single-trial time-varying inferred inputs u_t^i, representing input from unobserved brain regions.

Notation. All time-indexed variables and parameters are also indexed by trial, though we omit trial indices for notational simplicity. We use $t_1{:}t_2$ to denote the inclusive sequence of integers $\{t_1, t_1 + 1, \ldots, t_2\}$. We use $W^i(x) := W^i x + b^i$ to denote an affine transformation with

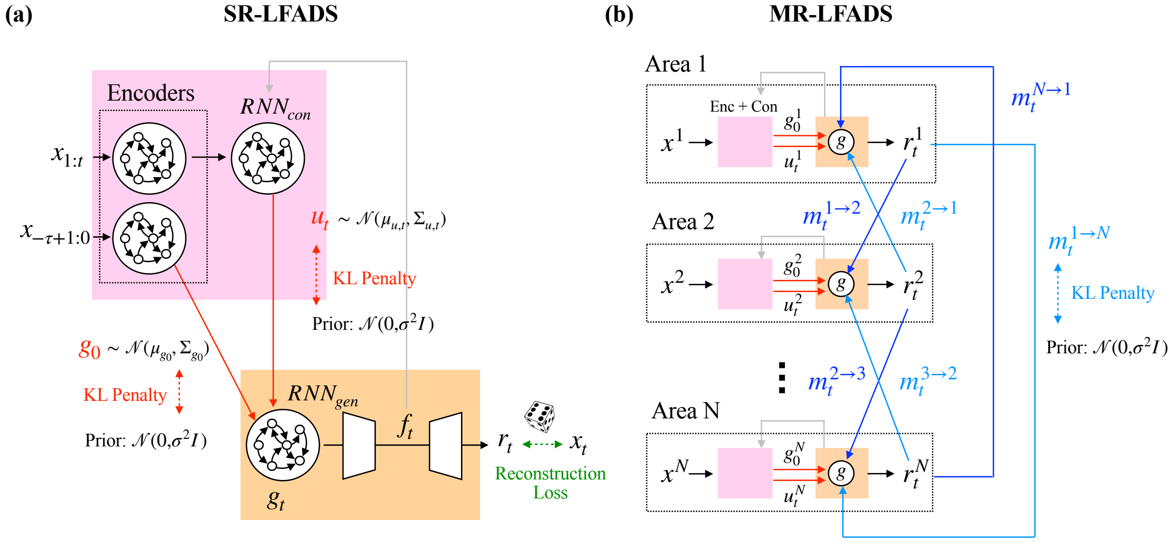

Figure 1. MR-LFADS architecture. (a) Adaptation of single-region LFADS with Poisson outputs. (b) MR-LFADS with N regions. KL penalties from SR-LFADS in panel (a) are replicated here, but are omitted in the diagram for clarity.

weights W^i and offsets b^i.

Generative Model. MR-LFADS treats all initial states g_0^i, communications $m_t^{j\to i}$, and inferred inputs u_t^i as latent variables. The prior distribution over these latent variables is modeled as:

$$g_0^i, u_t^i, m_t^{j\to i} \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

Given these quantities, the neural population dynamics in region i are modeled by a "generator" gated recurrent unit **(GRU)** network, $\text{GRU}^i_{\text{gen}}$, with internal states g_t^i that evolve according to:

$$g_t^i = \text{GRU}^i_{\text{gen}}\left(g_{t-1}^i, \left[\{m_{t-1}^{j\to i}\}_{j\neq i}; u_t^i\right]\right) \tag{2}$$

A set of region-i factors f_t^i are defined as an affine readout from the corresponding generator RNN:

$$f_t^i = W_f^i(g_t^i) \tag{3}$$

These factors are then transformed into parameters of time-varying output distributions in a manner dependent on the nature of the neural recordings. For continuous-valued observations, such as those from calcium imaging, a Gaussian or zero-inflated Gamma distribution may be appropriate (Zhu et al., 2022). In the Section 4 synthetic-data experiments, we apply a Gaussian output distribution:

$$P(x_t^i \mid g_0^i, u_{1:t}^i, \{m_{1:t}^{j\to i}\}_{j\neq i}) = \mathcal{N}(r_t^i, \Sigma_{r,t}^i) \tag{4}$$

$$r_t^i = W_r^i(f_t^i) \qquad \Sigma_{r,t}^i = \text{diag}\left(\exp(W_{\sigma_r}^i(f_t^i))\right) \tag{5}$$

where r_t^i and $\Sigma_{r,t}^i$ are the region-i predicted mean and covariance of the recorded neural activity, respectively, and are each computed via separate affine transformations, W_r^i and $W_{\sigma_r}^i$. For spike count observations, as modeled in the electrophysiology experiments of Section 5, we apply a Poisson output distribution:

$$P(x_t^i \mid \cdot) = \text{Poisson}(r_t^i) \qquad r_t^i = \exp\left(W_r^i(f_t^i)\right) \tag{6}$$

where the exponential nonlinearity ensures non-negative predicted firing rates r_t^i.

Inference Model. Following VAE conventions (Kingma & Welling, 2013), MR-LFADS approximates the intractable true posterior distributions over the latent variables using variational posteriors, which we denote as $q(\cdot|\cdot)$.

MR-LFADS defines the approximate posteriors over communication messages from observed region j to observed region i as Gaussian distributions with parameters derived from the region-j predicted firing rates r_t^j:

$$q(m_t^{j\to i} \mid x_{1:t}^j) = q(m_t^{j\to i} \mid r_t^j) = \mathcal{N}(\mu_{m,t}^{j\to i}, \Sigma_{m,t}^{j\to i}) \tag{7}$$

$$\begin{aligned} \mu_{m,t}^{j\to i} &= W_{\mu_m}^{j\to i}(r_t^j) \\ \Sigma_{m,t}^{j\to i} &= \text{diag}\Big(\exp\left(W_{\sigma_m}^{j\to i}(r_t^j)\right)\Big) \end{aligned} \tag{8}$$

Constraining communication to be derived from r_t^i anchors it to the neural recordings and in doing so reduces ambiguity in system identification. We refer to this rate-based communication model as **MR-LFADS(R)**. In Section 4, we explore generator-based **MR-LFADS(G)** and factor-based **MR-LFADS(F)** communication models, which replace all instances of r_t^j in Eq. 8 with g_t^j and f_t^j, respectively.

Approximate posteriors over the region-i initial generator states g_0^i and inferred inputs (from unobserved brain regions) u_t^i are defined as the following Gaussian distributions:

$$q(g_0^i \mid x_{-\tau:0}^i) = \mathcal{N}(\mu_{g_0}^i, \Sigma_{g_0}^i) \tag{9}$$

$$q(u_t^i \mid x_{1:t}^i) = \mathcal{N}(\mu_{u,t}^i, \Sigma_{u,t}^i) \tag{10}$$

where the mean and covariance parameters are computed from the corresponding conditioning recorded neural activity x^i via a set of region-i-specific "encoder" and "controller" GRU networks (see Appendix A.1). Our approach here slightly modifies the original SR-LFADS specification, which allows acausal inference via a bidirectional encoder network that processes each entire T-timestep neural recording $x_{1:T}$ to infer g_0 and each element of $u_{1:T}$. In contrast, we infer g_0^i (Eq. 9) using a bidirectional encoder applied only to past neural activity $x^i_{-\tau:0}$, preserving causality, and infer u_t^i (Eq. 10) using a unidirectional encoder RNN that processes $x^i_{1:t}$ in a strictly forward, causal manner. This formulation ensures that all predicted firing rates r_t^i, and thus all derived communication signals $m_t^{j\to i}$, are inferred causally from neural activity recorded up to time t.

Model Fitting. Following VAE conventions, MR-LFADS is trained by maximizing the evidence lower bound (**ELBO**), a variational lower bound on the data log-likelihood. The ELBO is a sum of two terms: (1) the reconstruction error

$$\sum_{t=1}^{T} \mathbb{E}_q\big[\log P(\{x_t^i\} \mid \{g_0^i\}, \{u_t^i\}, \{m_t^{j\to i}\})\big] \qquad (11)$$

and (2) the negative Kullback–Leibler (**KL**) divergence D_{KL} between the approximate posteriors (Eqs. 7–10) and the priors (Eq. 1) over the latent variables.

The reconstruction term is estimated by running samples from the approximate posteriors (Eqs. 7–10) through the generative model to evaluate the experiment-dependent output distributions from Eqs. 4–6. The D_{KL} term acts as a regularizing information bottleneck on the latent variables. To control this regularization, we allow rescaling of the D_{KL} term (Higgins et al., 2017; Keshtkaran et al., 2022). Noting that the D_{KL} term decomposes into contributions from the three sets of MR-LFADS latent variables $\{g_0^i\}$, $\{u_t^i\}$, and $\{m_t^{j\to i}\}$, we weight each contribution differently and treat the weights $(\beta_{g_0}, \beta_u, \beta_m)$ as hyperparameters. To encourage MR-LFADS to infer inputs from unobserved regions only when that information cannot be obtained as communication from an observed region, we propose a structured KL bottleneck with $\beta_u = 10\beta_m$. Other choices of KL regularization structure might be appropriate if *a priori* knowledge is available about information flow or anatomical connectivity between recorded regions. See Appendix A.1 for further detail on MR-LFADS.

We will release a PyTorch implementation of MR-LFADS upon the post-conference update to this paper.

4. Results I: Synthetic Multi-Region Datasets

Here, we evaluated MR-LFADS across a broad range of synthetic multi-region datasets (Experiments 1-3) that enable direct comparisons between ground truth and model-inferred communication.

Preliminaries. Each dataset was generated by a unique data-generating network (**DGN**): an ensemble of noisy RNN modules that are jointly trained to perform a specified cognitive neuroscience task. Each module represents an arbitrary brain region. Prior to training, we specified the presence or absence of a directed, low-rank communication channel between each pair of regions. In Experiments 1 and 2, we manually designed the DGNs to impose specific challenges outlined in Section 1. Experiment 3 evaluates MR-LFADS on data from dozens of randomly generated DGNs, each trained to perform a randomly selected cognitive neuroscience task. For all experiments, we treat the hidden-unit activity in each DGN module as recorded neural activity and maintain external inputs, inter-module connectivity, and communication between modules as ground truth. These ground truth quantities are crucial for evaluating communication models but are typically not directly observable in real-data settings.

We evaluate MR-LFADS on these datasets and compare MR-LFADS to a collection of existing communication models: RRR (Semedo et al., 2019); multi-population sticky recurrent SLDS (**mp-srSLDS**) (Glaser et al., 2020); and multi-region switching dynamical systems (**MR-SDS**) (Karniol-Tambour et al., 2024). We also compare against ablated variants of MR-LFADS to probe the value of specific MR-LFADS design features. RRR uses reduced-rank regression to predict each timestep of neural activity in one recorded region from the corresponding timestep of activity recorded in one or more other recorded regions. Like MR-LFADS, mp-srSLDS and MR-SDS are coupled sets of dynamical systems, with each dynamical system representing one recorded region. In mp-srSLDS each is implemented as an SLDS, and in MR-SDS each is implemented as a switching nonlinear dynamical system.

We evaluate communication models against a high bar: system identification of a causal model of each DGN from its multi-region dataset, including both the *pathways* and *content* of communication. To assess communication *pathways*, we consider recovery of an "effectome" (Pospisil et al., 2024) describing the causal flow of effects along the inter-regional connectome. We represent this effectome as a matrix where each element $(i, j)_{i\neq j}$ indicates the volume of directed communication flow from region j to i. The effectome reflects both the inter-regional connectivity and the extent to which communication flows over each directed connection. For each dataset, we compare model-inferred effectomes to the ground truth effectome by vectorizing the effectome matrices, removing elements corresponding to the matrix diagonal $(i = j)$, and computing cosine similarity $S_{\cos}$ between these two vectors. See Appendix B.1 for further detail.

To assess communication *content*, we evaluate the timestep-by-timestep correspondence between model-inferred and ground truth messages. To assess whether the inferred messages encode the information contained in the ground truth messages, we applied linear regression to predict the ground truth messages $m_t^{j\to i}$ from the inferred messages $\mu_{m,t}^{j\to i}$ (Eq. 7). We then report the R^2 score of the linear regression, denoted $R^2(\mu_m^{j\to i}, m^{j\to i})$. For MR-LFADS, we use an analogous procedure to compare model-inferred inputs $\mu_{u,t}^i$ (Eq. 10) to ground truth external inputs. This comparison is not possible for RRR, mp-srSLDS, or MR-SDS—these models do not infer inputs from unobserved regions. See Appendix B-D for details on each experiment, model hyperparameters, and evaluation metrics.

4.1. Experiment 1: Specifying Inputs.

This experiment demonstrates that manually specifying inputs in a model risks imposing erroneous communication structure, which can lead to inferring an effectome that misrepresents the true inter-regional communication. To evaluate the implications of input specification, we designed a DGN that implements a dynamical memory function. Each region of this "memory network" receives unique stimulus information from one unobserved region and from one observed region, and is tasked with remembering a recent history of those signals (Fig. 2a, left). This setup results in each region of the DGN representing information from two different sources, posing the challenge of disentangling whether each signal arises due to communication or due to external input. A common modeling choice is to provide all known external inputs to all model regions and to let model fitting determine which inputs are needed by each region. However, in this case, such manual input specification can result in a model completely forgoing communication (Fig. 2a, right) because the manually specified inputs contain the information that was actually transmitted as communication in the DGN.

We evaluate MR-LFADS(R), which does not use the stimulus signals during training or evaluation but rather automatically infers external inputs in an unsupervised manner. We also evaluate an MR-LFADS variant that ablates this inferred inputs feature. Termed MR-LFADS(S), this model receives all stimulus signals as manually specified external inputs to each region's generator $\text{GRU}_{\text{gen}}^i$ (replacing u_t^i with $s_t^{1:3}$ in Eq. 2). We also compare against MR-SDS, mp-srSLDS, and RRR.

The MR-LFADS variants (R and S) reconstructed the simulated multi-region activity more accurately than MR-SDS, mp-rSLDS, and RRR (Fig. 2b, Fig. S1a). Critically, MR-LFADS(R) accurately infers the effectome (Fig. 2c, left, Fig. S1b). By contrast, all models that are manually provided stimulus signals (MR-LFADS(S), MR-SDS, mp-rSLDS) infer less accurate effectomes and demonstrate the failure mode mentioned above, forgoing communication and instead relying on the specified inputs to provide the corresponding signals. RRR, which receives neither manually specified nor inferred inputs, also infers a less accurate effectome. These results suggest that manually specifying inputs can discourage models from utilizing—and thus identifying—communication. By automatically inferring inputs, MR-LFADS(R) avoids this failure mode.

Next, we assess the accuracy of inferred inputs and mes-

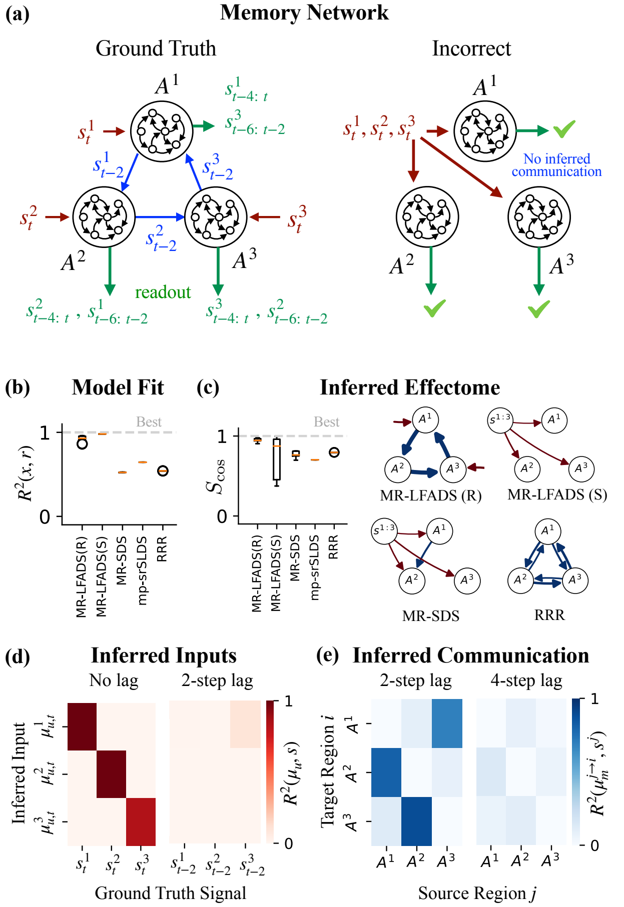

Figure 2. Memory network experiment. (a) *Left:* In this DGN, each region (*area*) A^i receives a private stimulus s_t^i (red) and communicates a two-step delayed version s_{t-2}^i (blue) to a downstream region. Each area is trained to recall the last five time steps of its private stimulus and its received communication (green). For example, readouts from A^2 are trained reproduce $s^2_{t-4:t}$, the private stimulus to A^2, and $s^1_{t-6:t-2}$, the private stimulus to A^2's upstream neighbor A^1. *Right:* Potential incorrect inferred communication model capable of reconstructing these synthetic data. (b) R^2 scores for data reconstruction. (c) *Left:* S_{cos} for inferred effectome. *Right:* Example fitted models, where arrow weights indicate the average message norm across trials and time. (d) R^2 of linear prediction of ground truth stimulus inputs (with time lag $\in \{0, 2\}$) from inferred inputs. (e) R^2 of linear prediction of ground truth messages (with time lag $\in \{2, 4\}$) via inferred messages.

sages. In the ground truth DGN at time t, region-i receives only s_t^i and $m_t^{j \to i}$. Thus, an accurate communication model should infer inputs and messages that encode only these time-t quantities. However, due to nature of the memory task, the DGN's region-i time-t RNN activities contain information about $s_{t-4:t}^i$ and $m_{t-4:t}^{j \to i} = s_{t-6:t-2}^j$. When fitting these data, a communication model is effectively posed with the question: how does that time-lagged information get into the region-i time-t simulated neural activity? MR-LFADS(R) correctly infers the current timestep ground truth private stimuli s_t^i as inputs to each region i (Fig. 2d, left). Importantly, MR-LFADS(R) also correctly avoids inferring time-lagged versions of those signals, which are inconsistent with the causal flow of information in the DGN, even though those signals are represented in the region-i activity and thus could support data reconstruction (Fig. 2d, right). Similarly, MR-LFADS(R) correctly infers the current timestep ground truth messages $m_t^{j \to i}$ (Fig. 2e, left) and avoids incorrectly inferring time-lagged versions, despite their utility toward data reconstruction (Fig. 2e, right). By contrast, all other models learn either no communication, or encode past information (Fig. S1c). See Appendix B.2 for further detail.

These results demonstrate the unique ability of MR-LFADS(R) to disentangle region-specific external inputs from communication between recorded regions, all in an unsupervised manner that mitigates biases associated with manual specification of external inputs. These results also imply that MR-LFADS(R) learns region-specific dynamics that transform those inputs and communications into region-specific representations that can reconstruct the data, thereby implementing a dynamical memory computation mimicking that of the DGN. See Appendix E.1 for further analyses demonstrating how the structured information bottlenecks robustly support this disentangling in MR-LFADS (Miller et al., 2024).

4.2. Experiment 2: Data-Constrained Communication

This experiment demonstrates the implications of message-inference design choices. We designed MR-LFADS(R) to infer messages as affine functions of the source-region predicted firing rates r_t^i, which themselves are tied to observed data x_t^i through the data reconstruction term in the ELBO. This architecture contrasts with that of mp-srSLDS and MR-SDS, which infer communication as a function of region-specific dynamical states. To directly evaluate these architectural implications, we compare the rate-based communication design of MR-LFADS(R) (Eqs. 7–8) to variants with factor-based and generator-based communication, termed MR-LFADS(F) and MR-LFADS(G), respectively.

To highlight the significance of this design choice, we evaluated models on data from a two-region "pass-decision" DGN that computes perceptual decisions based on time-varying sensory evidence (Fig. 3a, left). An upstream area A^P receives a white noise stimulus s_t and is trained to *pass* that stimulus through to a readout, effectively learning an identity function routing input to output. A downstream *decision* area A^D receives this stimulus as communication $m_t^{P \to D}$, integrates that stimulus over time into a decision variable d_t, and reports $^\pm 1$ choices indicating the sign of that decision variable (Mante et al., 2013).

This setup again poses the challenge of disentangling external inputs, communication, and local dynamics—and in particular, accurately identifying and localizing each ground truth dynamic. Though the integration dynamics are localized to A^D in the DGN, a sufficiently expressive communication model could incorrectly learn integration dynamics in A^P, while still accurately reconstructing the data, for example, if A^P represents s_t, integrates s_t into d_t, and communicates d_t as $m^{P \to D}$ to A^D (Fig. 3a, right).

All MR-LFADS models achieved comparable data recon-

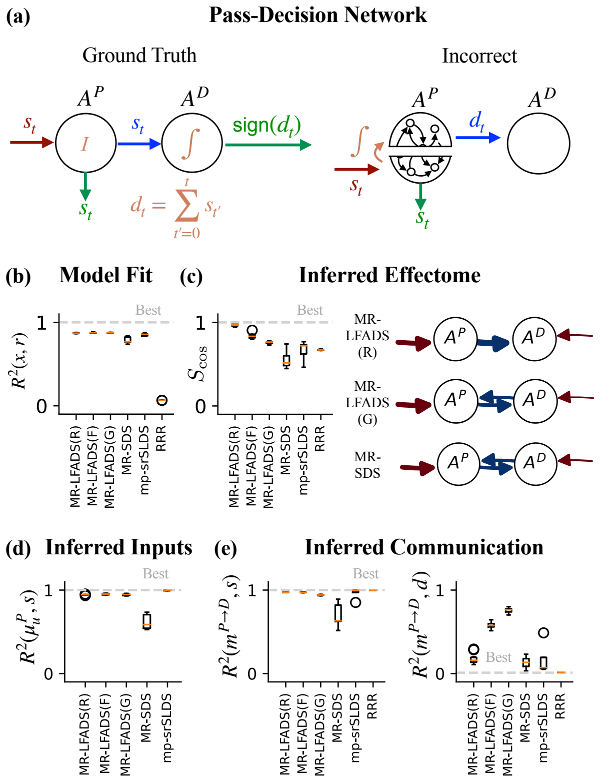

Figure 3. Pass-decision network experiment. (a) *Left:* DGN setup, with private stimulus (red), communication (blue), trained readouts (green), identity function (I), and integration function ($\int$). *Right:* Potential failure mode for a learned model. (b) R^2 scores for data reconstruction. (c) *Left:* S_{cos} for inferred effectome. *Right:* Example fitted models. (d) R^2 of linear prediction of ground truth input s to region P, from inferred inputs. (e) *Left:* R^2 of linear prediction of ground truth messages $m^{P \to D} = s$ from inferred messages, $\mu_m^{P \to D}$. *Right:* R^2 of linear prediction of the decision variable d from $\mu_m^{P \to D}$, indicating mislocalization of integration.

struction, slightly outperforming MR-SDS and mp-srSLDS (Fig. 3b, Fig. S2a). RRR reconstructed the data poorly, likely due to the dynamic nature of the integration computation, i.e., the decision variable d_t cannot be predicted from any single-timestep stimulus value $s_{t'}$.

Crucially, MR-LFADS(R) accurately inferred the effectome (Fig. 3c, left, Fig. S2b), identifying communication from A^P to A^D and not from A^D to A^P (Fig. 3c, right and Fig. S2b). By contrast, all other models incorrectly identified A^D to A^P communication. Notably, both MR-LFADS(F) and MR-LFADS(G) misidentified $A^D \rightarrow A^P$ communication, indicating that rate-based communication constraints in MR-LFADS(R) help mitigate excessive flexibility in expressive models.

The ground truth input to A^P is the stimulus s_t. All MR-LFADS models correctly inferred inputs to A^P that encode s_t (Fig. 3d, left). While MR-SDS and mp-srSLDS do not infer inputs in an unsupervised manner as in MR-LFADS, we can quantify the effective inputs they infer as a function of their manually specified inputs and the corresponding trained input mappings. MR-SDS inputs to A^P carried markedly less information about s_t relative to MR-LFADS and mp-srSLDS.

Accurate identification of communication requires inferred messages $m_t^{P\rightarrow D}$ to encode the stimulus s_t. By contrast, $m_t^{P\rightarrow D}$ instead encoding the decision variable d_t would imply mislocalization of the integration dynamic (Fig. 3a, right). Only MR-LFADS(R), mp-srSLDS, and RRR correctly encoded s_t in $m_t^{P\rightarrow D}$ without incorrectly encoding d_t (Fig. 3e, right). That MR-LFADS(R) succeeds, whereas MR-LFADS(F) and MR-LFADS(G) both misidentify $A^P \rightarrow A^D$ communication as encoding d_t, again demonstrates the importance of data-constrained communication for mitigating the risks of excessive flexibility in expressive models. See Appendix E.2 for further analyses into this excessive flexibility.

4.3. Experiment 3: Random Multi-Region Networks

The previous experiments were designed to highlight specific failure modes of communication models. However, real brain-wide networks exhibit a broad range of architectures and dynamics. To assess generalization across such a broad range of settings, here we evaluate communication models on datasets generated by a wide variety of randomly configured multi-region DGNs (Fig. 4a, top), each trained to perform a randomly selected cognitive neuroscience task (Fig. 4a, bottom).

We generated 35 multi-region datasets, each from a unique

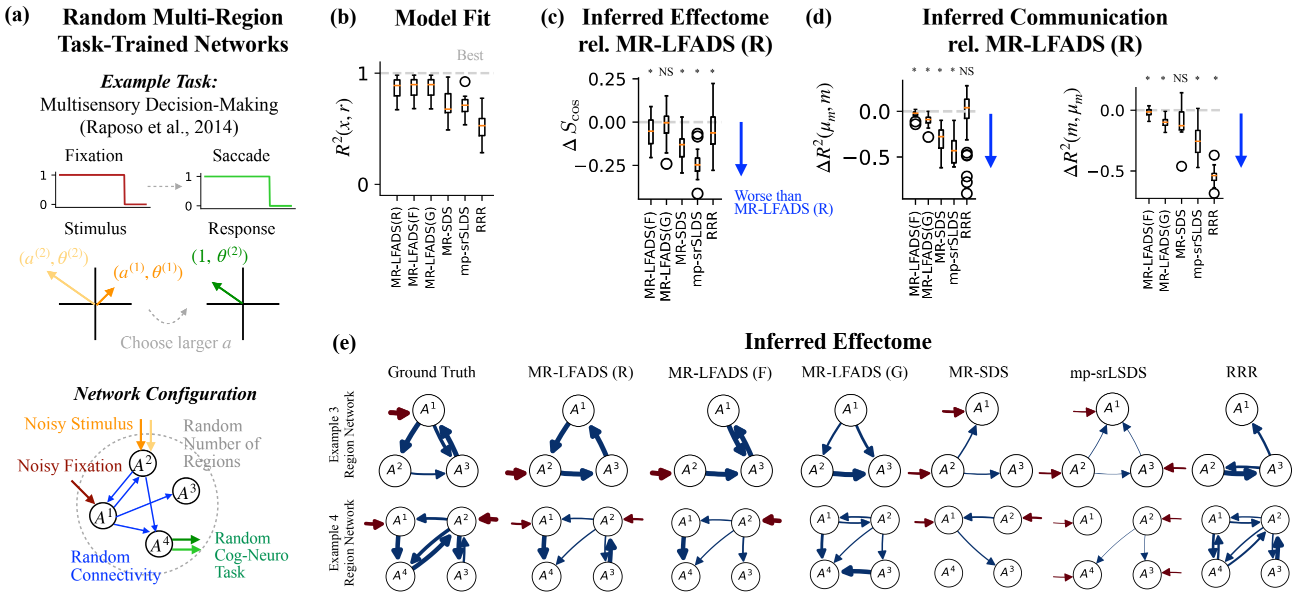

Figure 4. Randomly generated multi-region network experiments. (a) *Top:* Multisensory decision-making as an example cognitive task. When the fixation cue disappears, the output area A^N must saccade and report the $\theta^{(i)}$ value corresponding to the stimulus with larger $a^{(i)}$. *Bottom:* Each DGN is configured with a random number of areas (3 or 4), random inter-regional connectivity, and is trained on a randomly selected task. (b) R^2 scores for data reconstruction. (c) $S_{\cos}$ for each specified model minus $S_{\cos}$ for MR-LFADS(R). One-tailed t-test p-values: $p = 0.00016, 0.12, 0.006, 0.0, 0.01$. (d) *Left:* Differences of R^2 scores, relative to MR-LFADS(R), for linear prediction of ground truth messages from inferred messages. One-tailed t-test p-values: $p = 0.0, 0.0, 0.001, 0.0, 0.14$. *Right:* Differences of R^2 scores for linear prediction of inferred messages from ground truth messages. One-tailed t-test p-values: $p = 0.0004, 0.0, 0.08, 0.0, 0.0$. (e) We selected networks with median $S_{\cos}$ scores for MR-LFADS(R), one with three areas (*top*) and another with four (*bottom*). Arrow thicknesses indicate relative message norms.

DGN. Each DGN consisted of three or four regions with randomly generated inter-regional connectivity. Tasks were drawn from the set described by Yang et al. (2019), spanning multiple variants of decision-making, working memory, categorization, and inhibitory control (see Appendix D). On each trial, the DGN received a noisy fixation stimulus $s_{\text{fix},t}$ and two noisy task stimuli, represented in polar coordinates as $s_t^k = (a_t^{(k)}, \theta_t^{(k)})$ for $k \in \{1, 2\}$. The DGN was trained to process these inputs and output a task-dependent response angle θ^{resp} and a task-dependent eye movement.

We fit MR-LFADS(R), (F), and (G), along with mp-srSLDS, MR-SDS, and RRR, to each of these datasets. Aggregating results across all datasets, the MR-LFADS models achieved the best data reconstruction, which was indistinguishable across model variants (Fig. 4b, Fig. S3). MR-LFADS(R) and MR-LFADS(G) inferred the most accurate effectomes, with statistically indistinguishable $S_{\cos}$ distributions (Fig. 4c). Next, we evaluated the accuracy of inferred message content. We predicted the ground truth messages from inferred messages and used R^2 scores as a measure of how much information the inferred messages contain about the ground truth (Fig. 4d, left). We also performed the reverse, predicting the inferred messages from the ground truth and interpreted lower R^2 scores as an indication that inferred messages contained additional information beyond that present in the ground truth (Fig. 4d, right). MR-LFADS(R) was the only model that performed best across both of these metrics. Inferred effectomes from two example datasets are shown in Fig. 4e. MR-LFADS(R) inferred effectomes most similar to the ground truth. However, it is not perfect and does miss a connection that is also not detected by the other models. Taken together, these results demonstrate that MR-LFADS(R) outperforms existing communication models across a broad range of neuroscience-relevant synthetic multi-region datasets.

5. Results II: Multi-Region Electrophysiology

Here, we applied MR-LFADS(R) to large-scale electrophysiology data from multiple simultaneously recorded Neuropixel probes in mice performing a decision-making task (Fig. 5a) (Chen et al., 2024). We evaluated how well MR-LFADS can predict experimentally observed effects of causal circuit perturbations that were not included in the data used to train the model. We also evaluated the reliability of inferred communication across random initializations of the model, comparing MR-LFADS(R) to MR-LFADS(G).

We fit a 5-region MR-LFADS model to population activity recorded across anterior lateral motor cortex (**ALM**), midbrain reticular nucleus (**MRN**), superior colliculus (**SC**), thalamic regions known to be strongly and reciprocally connected with ALM (**Thal(A)**), and other thalamic regions (**Thal(O)**). In a subset of trials, ALM was briefly photoinhibited (Fig. 5a). MR-LFADS was trained only on unperturbed ("control") trials, with photoinhibition trials held out for post-training validation. See Appendix F for further detail.

We first confirmed that MR-LFADS accurately fits held-out control trials (Fig. 5b). Next, we sought to adapt MR-LFADS(R) to predict experimentally observed effects of ALM photoinhibition (Fig. 5b, c). To mimic ALM pho-

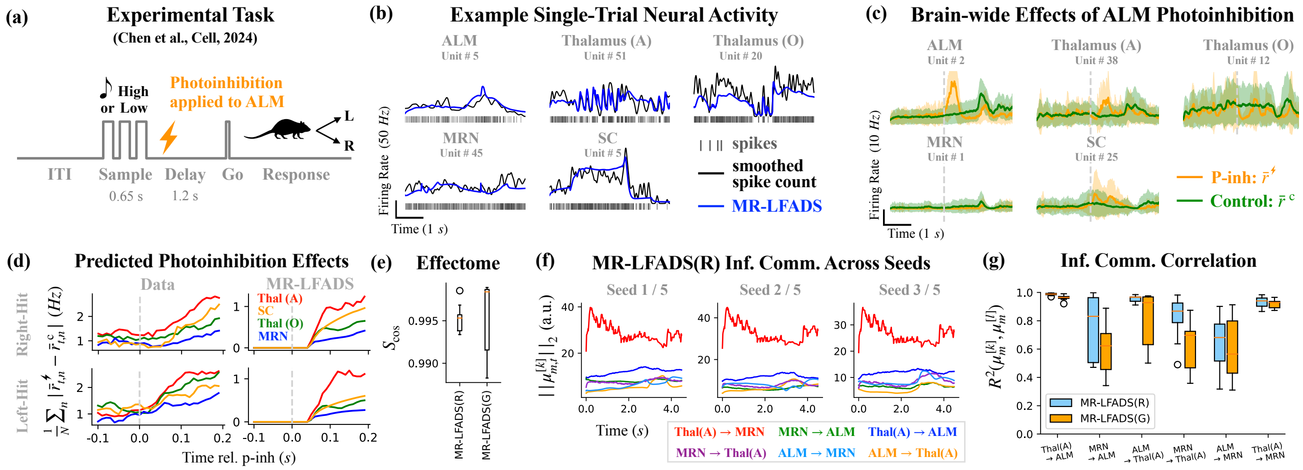

Figure 5. MR-LFADS(R) applied to multi-region, high-density electrophysiology recordings. (a) Mice receive a high- or low-tone auditory stimulus (sample period) and respond by licking left or right (response period). (b) MR-LFADS single-trial predicted firing rates in held-out control trials (blue), recorded spike times (black vertical ticks), and smoothed, binned spike counts (black; causal, exponential filter) for example neurons in each modeled brain region. (c) Condition-averaged smoothed spike counts of example neurons in control and photoinhibition trials (left-hit condition). Shaded regions indicate the standard deviation across trials. (d) Photoinhibition-related changes in population recordings, as observed experimentally (left) and as predicted by MR-LFADS (right). (e) Cosine similarity of inferred effectomes across models with different random initializations (seeds). (f) Message norms inferred by MR-LFADS(R) for all connections across example seeds k. (g) Correlation of inferred message norms across pairs of seeds (k, l).

toinhibition *in silico*, we ablated MR-LFADS communication from ALM to the other regions by setting those messages to 0. We summarized the temporal influence of ALM photoinhibition by computing the differences between condition-averaged population activity in photoinhibition ($\bar{r}_t^{\lightning}$) and control ($\bar{r}_t^{c}$) trials. MR-LFADS predicted these photoinhibition effects (Fig. 5d), despite never seeing photoinhibition trials during training. Namely, MR-LFADS predicted that Thal(A) should be most affected by ALM photoinhibition, MRN least affected, and SC and Thal(O) intermediately affected. These results demonstrate that MR-LFADS learned a model of inter-regional communication that is accurate enough to predict effects of causal multi-region circuit perturbations.

Finally, we sought to evaluate the consistency of MR-LFADS models across random initializations of the model parameters. We compared MR-LFADS(R) and MR-LFADS(G), each trained from five random seeds on simultaneous population recordings from ALM, thalamus, and MRN (see Appendix F2). Both models inferred consistent effectomes (Fig. 5e), but MR-LFADS(R) was more consistent in terms of inferred message content (Fig. 5f, g), further supporting the utility of data-constrained communication—particularly for improving the reproducibility of scientific conclusions derived from the model.

6. Discussion

Understanding how brain regions interact to support distributed computation requires communication models that can disentangle inter-regional communication from local population dynamics and inputs originating from unrecorded regions. In this work, we identified critical failure modes that limit existing communication models—including misidentification due to manually specified external inputs and mislocalization of neural dynamics. We introduced MR-LFADS, a communication model specifically designed to mitigate such failures through three key design features: (1) automatic inference of region-specific inputs from unobserved sources, (2) communication constrained to originate from data-linked reconstructed firing rates, and (3) structured regularization that promotes disentangling and prevents inferring inputs from unobserved sources when the same information can be obtained via communication from observed regions. These features help prevent the model from learning spurious solutions that fit the data well but yield misleading conclusions about network interactions.

Using synthetic datasets engineered to pressure-test communication models, we demonstrated that MR-LFADS outperforms existing approaches in accurately recovering both the structure and content of inter-regional communication. Crucially, ablated MR-LFADS variants demonstrated that these performance gains depend critically on specific MR-LFADS design features. These results generalized beyond carefully engineered scenarios, with MR-LFADS outperforming existing models also across a broad range of multi-region datasets synthesized from a distribution of task-trained multi-region networks.

Applying MR-LFADS to real multi-region electrophysiological recordings further validated its utility, as the model inferred inter-regional interactions that accurately predicted brain-wide effects of causal perturbations, despite these perturbations being absent during training. Here, MR-LFADS models were also more reproducible across random model initializations compared to a model variant that removed data-constraints on inferred communication, suggesting that this design feature specifically enhances the reproducibility of model-derived scientific conclusions.

Despite its advantages, MR-LFADS has potential limitations. It remains unclear whether the model will succeed in scenarios in which its communication-favoring inductive bias is not appropriate—for example, when a region's dynamics are largely autonomous, are only weakly influenced by other recorded regions, or are influenced by an unobserved region whose activity is correlated with that in a non-communicating observed region. Additionally, MR-LFADS remains sensitive to hyperparameter settings, mirroring a known limitation of SR-LFADS (Keshtkaran et al., 2022). Thus, significant computational resources might be required to adequately optimize hyperparameters. Finally, as in SR-LFADS, hyperparameters specifying the prior distributions in MR-LFADS can shape the representations in inferred latent variables, which can in turn shape the region-specific dynamics learned to reconstruct the neural recordings. While we have reason to believe that such inferred quantities can indicate the presence, identity, and timing of inputs (Pandarinath et al., 2018) and communications, future work is needed to interpret the representations of those signals (Sedler et al., 2023; Versteeg et al., 2024).

Acknowledgments

This work was supported by NIH award T32-MH132518 (JS), the Paul G. Allen Foundation (MDG), and NIH award R00-MH121533 (MDG). We thank O. Karniol-Tambour for valuable discussion and support with MR-SDS (Karniol-Tambour et al., 2024). We are grateful to S. Chen, N. Steinmetz, E. Shea-Brown, and J.N. Kutz for insightful input on the project, and to H. Gurnani, T. Kim, and J. Pemberton for insightful feedback on the manuscript. We are also grateful to D. Sussillo, C. Pandarinath, W. Newsome, and K. Shenoy for many years of guidance and inspiration relevant to this project.

Impact Statement

This work aims to advance the fields of Machine Learning and Neuroscience, with potential societal impacts including the development of neuroengineering technologies and treatments for neurological injuries, diseases, and neuropsychiatric conditions.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Allen, W. E., Kauvar, I. V., Chen, M. Z., Richman, E. B., Yang, S. J., Chan, K., Gradinaru, V., Deverman, B. E., Luo, L., and Deisseroth, K. Global representations of goal-directed behavior in distinct cell types of mouse neocortex. *Neuron*, 94(4):891–907, 2017.

Allen, W. E., Chen, M. Z., Pichamoorthy, N., Tien, R. H., Pachitariu, M., Luo, L., and Deisseroth, K. Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science*, 364(6437):eaav3932, 2019.

Bennett, C., Ouellette, B., Ramirez, T. K., Cahoon, A., Cabasco, H., Browning, Y., Lakunina, A., Lynch, G. F., McBride, E. G., Belski, H., et al. Shield: Skull-shaped hemispheric implants enabling large-scale electrophysiology datasets in the mouse brain. *Neuron*, 112(17): 2869–2885, 2024.

Biswas, T., Bishop, W. E., and Fitzgerald, J. E. Theoretical principles for illuminating sensorimotor processing with brain-wide neuronal recordings. *Current Opinion in Neurobiology*, 65:138–145, 2020.

Chen, S., Liu, Y., Wang, Z. A., Colonell, J., Liu, L. D., Hou, H., Tien, N.-W., Wang, T., Harris, T., Druckmann, S., et al. Brain-wide neural activity underlying memory-guided movement. *Cell*, 187(3):676–691, 2024.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.

Duncker, L. and Sahani, M. Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. *Current Opinion in Neurobiology*, 70:163–170, 2021.

Durstewitz, D., Koppe, G., and Thurm, M. I. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nature Reviews Neuroscience*, 24(11):693–710, 2023.

Gilad, A., Gallero-Salas, Y., Groos, D., and Helmchen, F. Behavioral strategy determines frontal or posterior location of short-term memory in neocortex. *Neuron*, 99 (4):814–828, 2018.

Glaser, J., Whiteway, M., Cunningham, J. P., Paninski, L., and Linderman, S. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in Neural Information Processing Systems*, 33: 14867–14878, 2020.

Gokcen, E., Jasper, A. I., Semedo, J. D., Zandvakili, A., Kohn, A., Machens, C. K., and Yu, B. M. Disentangling the flow of signals between populations of neurons. *Nature Computational Science*, 2(8):512–525, 2022.

Gokcen, E., Jasper, A., Xu, A., Kohn, A., Machens, C. K., and Yu, B. M. Uncovering motifs of concurrent signaling across multiple neuronal populations. *Advances in Neural Information Processing Systems*, 36, 2024.

Goris, R. L., Movshon, J. A., and Simoncelli, E. P. Partitioning neuronal variability. *Nature Neuroscience*, 17(6): 858–865, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

IBL, Benson, B., Benson, J., Birman, D., Bonacchi, N., Bougrova, K., Bruijns, S. A., Carandini, M., Catarino, J. A., Chapuis, G. A., et al. A brain-wide map of neural activity during complex behaviour. *bioRxiv preprint*, 2023.

Jia, X., Siegle, J. H., Durand, S., Heller, G., Ramirez, T. K., Koch, C., and Olsen, S. R. Multi-regional module-based signal transmission in mouse visual cortex. *Neuron*, 110 (9):1585–1598, 2022.

Kang, B. and Druckmann, S. Approaches to inferring multi-regional interactions from simultaneous population recordings. *Current Opinion in Neurobiology*, 65: 108–119, 2020.

Karniol-Tambour, O., Zoltowski, D. M., Diamanti, E. M., Pinto, L., Brody, C. D., Tank, D. W., and Pillow, J. W. Modeling state-dependent communication between brain regions with switching nonlinear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024.

Kass, R. E., Bong, H., Olarinre, M., Xin, Q., and Urban, K. N. Identification of interacting neural populations: methods and statistical considerations. *Journal of Neurophysiology*, 130(3):475–496, 2023.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., and Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience*, 17(3): 440–448, 2014.

Keeley, S. L., Zoltowski, D. M., Aoi, M. C., and Pillow, J. W. Modeling statistical dependencies in multi-region spike train data. *Current Opinion in Neurobiology*, 65: 194–202, 2020.

Keshtkaran, M. R., Sedler, A. R., Chowdhury, R. H., Tandon, R., Basrai, D., Nguyen, S. L., Sohn, H., Jazayeri, M., Miller, L. E., and Pandarinath, C. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *Nature Methods*, 19 (12):1572–1577, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint*, 1312.6114, 2013.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Linderman, S. W., Miller, A. C., Adams, R. P., Blei, D. M., Paninski, L., and Johnson, M. J. Recurrent switching linear dynamical systems. *arXiv preprint*, 1610.08466, 2016.

MacDowell, C. J., Libby, A., Jahn, C. I., Tafazoli, S., Ardalan, A., and Buschman, T. J. Multiplexed subspaces route neural activity across brain-wide networks. *Nature Communications*, 16(1):3359, 2025.

Makino, H., Ren, C., Liu, H., Kim, A. N., Kondapaneni, N., Liu, X., Kuzum, D., and Komiyama, T. Transformation of cortex-wide emergent properties during motor learning. *Neuron*, 94(4):880–890, 2017.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

Miller, K., Eckstein, M., Botvinick, M., and Kurth-Nelson, Z. Cognitive model discovery via disentangled rnns. *Advances in Neural Information Processing Systems*, 36, 2024.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., and Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22(10):1677–1686, 2019.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., et al. Inferring single-trial neural population dynamics using sequential autoencoders. *Nature Methods*, 15(10):805–815, 2018.

Perich, M. G. and Rajan, K. Rethinking brain-wide interactions through multi-region 'network of networks' models. *Current Opinion in Neurobiology*, 65:146–151, 2020.

Perich, M. G., Arlt, C., Soares, S., Young, M. E., Mosher, C. P., Minxha, J., Carter, E., Rutishauser, U., Rudebeck, P. H., Harvey, C. D., et al. Inferring brain-wide interactions using data-constrained recurrent neural network models. *bioRxiv preprint*, 2020.

Pospisil, D. A., Aragon, M. J., Dorkenwald, S., Matsliah, A., Sterling, A. R., Schlegel, P., Yu, S.-c., McKellar, C. E., Costa, M., Eichler, K., et al. The fly connectome reveals a path to the effectome. *Nature*, 634(8032):201–209, 2024.

Raposo, D., Kaufman, M. T., and Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17(12):1784–1792, 2014.

Ruff, D. A. and Cohen, M. R. Simultaneous multi-area recordings suggest that attention improves performance by reshaping stimulus representations. *Nature Neuroscience*, 22(10):1669–1676, 2019.

Sedler, A. R., Versteeg, C., and Pandarinath, C. Expressive architectures enhance interpretability of dynamics-based neural population models. *Neurons, Behavior, Data Analysis, and Theory*, 2023, 2023.

Semedo, J. D., Zandvakili, A., Machens, C. K., Byron, M. Y., and Kohn, A. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259, 2019.

Semedo, J. D., Gokcen, E., Machens, C. K., Kohn, A., and Byron, M. Y. Statistical methods for dissecting interactions between brain areas. *Current Opinion in Neurobiology*, 65:59–69, 2020.

Shenoy, K. V., Sahani, M., and Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annual Review of Neuroscience*, 36(1):337–359, 2013.

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

Sofroniew, N. J., Flickinger, D., King, J., and Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife*, 5:e14472, 2016.

Song, A., Charles, A. S., Koay, S. A., Gauthier, J. L., Thiberge, S. Y., Pillow, J. W., and Tank, D. W. Volumetric two-photon imaging of neurons using stereoscopy (vTwINS). *Nature Methods*, 14(4):420–426, 2017.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., and Harris, K. D. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364 (6437):eaav7893, 2019.

Versteeg, C., Sedler, A. R., McCart, J. D., and Pandarinath, C. Expressive dynamics models with nonlinear injective readouts enable reliable recovery of latent features from neural activity. In *Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 228 of *Proceedings of Machine Learning Research*, pp. 255–278. PMLR, 2024.

Veuthey, T., Derosier, K., Kondapavulur, S., and Ganguly, K. Single-trial cross-area neural population dynamics during long-term skill learning. *Nature Communications*, 11(1):4057, 2020.

Vyas, S., Golub, M. D., Sussillo, D., and Shenoy, K. V. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43(1):249–275, 2020.

Watanabe, S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S., Shenoy, K. V., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in Neural Information Processing Systems*, 21, 2008.

Zhu, F., Grier, H. A., Tandon, R., Cai, C., Agarwal, A., Giovannucci, A., Kaufman, M. T., and Pandarinath, C. A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with subframe temporal resolution. *Nature Neuroscience*, 25 (12):1724–1734, 2022.

A. Models

A.1. Multi-Region LFADS (MR-LFADS)

Reconstruction validation. To evaluate model fit, we assess how well MR-LFADS reconstructs the activity of 10% of neurons held out during training. A separate linear decoder, W_{hout}, is trained *post hoc* to predict held-out activity x_{hout} from the inferred latent factors. Importantly, this decoder is not trained end-to-end with MR-LFADS, ensuring that held-out neurons do not influence model learning. Fit quality is quantified using the R^2 score between predicted and actual held-out activity.

KL penalty. A critical hyperparameter during training is the scale of the KL penalty. The KL penalty coefficient for the inferred inputs, β_u, is always set higher than that for communication, β_m, an implicit assumption that encourages the network to prioritize learning information via communication channels whenever possible. The timing of when KL penalties are introduced also impacts results, though to a lesser extent. The schedule we found to work well begins with an initial stage where no KL penalty is applied, allowing the model to overfit to the data. Next, the penalty for inferred inputs is introduced, discouraging the model's reliance on these inputs. Finally, the penalty for communication is added, limiting the model from learning excessive information through communication channels.

Weight regularization. We apply light L_2 regularization to all GRU network recurrent weights.

Hyperparameters. Table S1 summarizes key hyperparameters used in the MR-LFADS models. Overall, we find that KL coefficients have the most impact on held-out neuron loss, S_{cos}, and R^2 scores for messages compared to other hyperparameters. SR-LFADS was originally described with a factor layer that is potentially lower dimensional than the number of generator units or modeled neurons. Here, we remove the rank constraint from the generator hidden states to rates by setting $N_{\text{fac}} = N_{\text{neu}}$. Additionally, the inferred input and message channel dimensions only need to exceed the estimated true dimensionality of these quantities, as KL penalties naturally suppress redundant channels by driving their activity to zero, as discussed in Appendix E.1.

Table S1. Key hyperparameters for MR-LFADS models. Experiment 4 refers to applications to multi-region electrophysiology data.

Hyperparameter	value	Description
learning rate	$\in [10^{-5}, 0.004]$	Scheduled by PyTorch's `ReduceLROnPlateau`; initial value: 0.004
T	190	Total time used for inferring inferred inputs
τ	10	Total time used for inferring the initial condition
total epoch	350	Total number of epochs
β_u		KL penalty coefficient for u; performs search for this hyperparameter
β_u start epoch	50	Epoch at which β_u starts increasing from 0
β_u increase epoch	200	Number of epochs for β_u to reach the maximum value
β_m		KL penalty coefficient for m; performs search for this hyperparameter
β_m start epoch	150	Epoch at which β_m starts increasing from 0
β_m increase epoch	100	Number of epochs for β_m to reach the maximum value
β_{g_0}	β_u	KL penalty coefficient for g_0
α	10^4	L2 penalty coefficient
α start epoch	0	Epoch at which α starts increasing from 0
α increase epoch	80	Number of epochs for α to reach the maximum value
N^i_{neu}		Number of neurons, (64, 16, 64) for exp 1, 2 and 3 respectively
N^i_{gen}	$2N^i_{\text{Neu}}$	Generator size
N^i_{fac}	N^i_{Neu}	Factor size
N^i_{inp}		Inferred input dimension, $(4, 8, 6, 8)$ for exp 1, 2, 3 and 4 respectively
N^i_{msg}		Inferred message dimension, $(4, 16, 10, 8)$ for exp 1, 2, 3 and 4 respectively

Unidirectional Encoder and Controller. To ensure that inferred inputs reflect only causal information—so that messages, in turn, are causal and thereby allow a more mechanistic interpretation in MR-LFADS—we modify the original LFADS model so that both the encoder and controller used for input inference are entirely unidirectional:

$$\begin{aligned} e_t^i &= \text{GRU}^i_{\text{enc},u}(e^i_{t-1}, x^i_t) \\ c_t^i &= \text{GRU}^i_{\text{con}}(c^i_{t-1}, [e^i_t; f^i_{t-1}]) \end{aligned} \tag{12}$$

The inferred inputs are then given by:

$$\begin{aligned} &q(u_t^i \mid x^i_{1:t}) = q(u^i_t \mid c^i_t) = \mathcal{N}(\mu^i_{u,t}, \Sigma^i_{u,t}) \\ &\mu^i_{u,t} = W^i_{\mu_u}(c^i_t) \quad \Sigma^i_{u,t} = \text{diag}\Big(\exp\big(W^i_{\sigma_u}(c^i_t)\big)\Big) \end{aligned} \tag{13}$$

By contrast, the encoder for the initial condition can remain bidirectional, since it operates on data preceding $t = 1$ and thus does not violate causality:

$$\begin{aligned} e_t^{i,-} &= \text{GRU}^{i,-}_{\text{enc},g}(e^{i,-}_{t+1}, x^i_t) \\ e_t^{i,+} &= \text{GRU}^{i,+}_{\text{enc},g}(e^{i,+}_{t-1}, x^i_t) \end{aligned} \tag{14}$$

$$\begin{aligned} q(g_0^i \mid x^i_{-\tau:0}) = q(g^i_0 \mid [e^{i,-}_{-\tau}; e^{i,+}_0]) &= \mathcal{N}(\mu^i_{g_0}, \Sigma^i_{g_0}) \\ \mu^i_{g_0} &= W^i_{\mu_{g_0}}([e^{i,-}_{-\tau}; e^{i,+}_0]) \\ \Sigma^i_{g_0} &= \text{diag}\Big(\exp\big(W^i_{\sigma_{g_0}}([e^{i,-}_{-\tau}; e^{i,+}_0])\big)\Big) \end{aligned} \tag{15}$$

A.2. Reduced-Rank Regression

The RRR model used in this study is based on the inter-regional communication subspace model of MacDowell et al. (2025), which integrates reduced-rank regression with ridge regression. The key difference in our implementation is that we apply the rank constraint separately to each source brain region, allowing us to disentangle the contributions of individual areas. Specifically, rather than concatenating activity from all regions into a single input matrix governed by a shared rank-constrained weight matrix—which conflates signals across regions—we assign a dedicated weight submatrix to each source region A^j, each with its own rank constraint r^j.

Therefore, for the communication subspace model, we have:

$$\begin{aligned} &W^{\text{rrr}} = \underset{\text{rank}(W)=r}{\arg\min} \left\| Y - XW \right\|_F^2 + \alpha \left\| W \right\|_F^2 \\ &\text{which is equivalent to:} \\ &W^{\text{ridge}} = \arg\min_W \left\| Y - XW \right\|_F^2 + \alpha \left\| W \right\|_F^2 \\ &W^{\text{rrr}} = \arg\min_W \left\| Y - XW^{\text{ridge}} \right\|_F^2 \\ &\qquad\qquad + \left\| XW^{\text{ridge}} - XW \right\|_F^2 \\ &\text{which is then equivalent to:} \\ &W^{\text{rrr}} = W^{\text{ridge}} V_r V_r^T, \text{where } U\Sigma V^T = XW^{\text{ridge}} \end{aligned} \tag{16}$$

where $X \in \mathbb{R}^{T \times N_{\text{src}}}$ represents the activity of all source regions concatenated together, and $Y \in \mathbb{R}^{T \times N_{\text{tar}}}$ represents the activity of the target region. W^{ridge} is the weight matrix obtained after applying ridge regression, and α is the ridge regularization parameter. W^{rrr} is the reduced-rank regression matrix obtained after ridge regression is applied. In the final step, singular value decomposition (SVD) is applied to XW^{ridge}, where $U\Sigma V^T$ is the decomposition, and V_r corresponds to the top r components of V.

In this version, the W^{ridge} matrix is divided into chunks corresponding to different regions A^j:

$$W^{\text{ridge}} = [W^{\text{ridge},1}; W^{\text{ridge},2}; \ldots; W^{\text{ridge},N}], \tag{17}$$

where each $W^{\text{ridge},j}$ corresponds to the contribution of source region A^j, and $[W^1; ...; W^N]$ represents a vertical stack of the matrices. SVD is then applied to each individual $W^{\text{ridge},j}$ matrix:

$$U^j \Sigma^j (V^j)^T = W^{\text{ridge},j}. \tag{18}$$

The reduced-rank version of $W^{\text{ridge},j}$ is computed as:

$$W^{\text{rrr},j} = W^{\text{ridge},j} V^j_{r^j} (V^j_{r^j})^T. \tag{19}$$

Finally, all $W^{\text{rrr},j}$ matrices are concatenated to form the complete W^{rrr} matrix. This ensures that the rank reduction applied to each $W^{\text{ridge},j}$ only compresses the information within that specific region's contribution, preserving the interpretability of the communication pathway from A^j to the target region.

The hyperparameters for this model include α and a matrix $R \in \mathbb{R}^{N \times N}$, where each element r^{ij} represents the rank associated with the communication from source region A^j to target region A^i. Additionally, since time delays may exist between regions—and such delays are explicitly configured in the synthetic datasets—a delay parameter d^{ij} is introduced for each source-target communication channel.

For fitting the memory and pass-decision networks, to tune the model, we perform an iterative search based on cross-validation performance, optimizing the hyperparameters in the following order: $D = \{d^{ij} : i, j = 1, \ldots, N, i \neq j\}$, α, and $R = \{r^{ij} : i, j = 1, \ldots, N, i \neq j\}$. This process is repeated until the hyperparameter values converge. The final values used are provided in Table S2. While the rank values R for both networks did not converge exactly to the true ranks of the messages, they were close. Notably, providing the true number of latents did not necessarily lead to better results. The delay values D were accurately learned for both models.

For fitting the networks in Experiment 3, the true delay is directly provided, and other hyperparameters are iterated in the same order for 10 epochs.

To increase the robustness of the RRR model fit, we implemented a bagging approach. For each model, 10 trials were bootstrapped from the training set, with each trial containing 200 time steps. A total of 87 fitted models were averaged to obtain the matrix W^{rrr}. This specific number of models was chosen to ensure that the total number of trials used during training remained consistent with other models.

Table S2. Hyperparameter search results for RRR models.

	Memory	Pass-Decision
α	0.055	0.01
R	$\begin{pmatrix} & 12 & 24 \\ 12 & & 32 \\ 24 & 18 & \end{pmatrix}$	$\begin{pmatrix} & 1 \\ 6 & \end{pmatrix}$
d^{ij}, $i \neq j$	2	0

A.3. Multi-Region Switching Dynamical Systems

We consider two variants of multi-region switching dynamical system models. The first is mp-srSLDS (Glaser et al., 2020), which consists of linear transitions, dynamics, and emissions. The relevant hyperparameters are the number of latent states per region, the number of discrete switching states, amount of L_1 and L_2 regularization on the weights, and the learning rate. Additionally, we consider MR-SDS (Karniol-Tambour et al., 2024), which is an extension that uses nonlinear transitions, dynamics, and emissions. It consists of two components: an inference network and a latent state-space model. The inference network is a transformer that performs the amortized inference of latent variables given observed neural activity. The latent state-space model is composed of a number of networks and functions as a structured prior on the latent variables. Specifically, we consider additive communication and input terms to the latent dynamics of the state-space model. That is, messages from other regions and external inputs affected the latent dynamics via additive terms. Relevant hyperparameters include the number of latent states per region, the number of discrete switching states, and the sizes of each sub-network.

For both models, we did extensive hyperparameter tuning to find the best model for each of the synthetic datasets and then computed all metrics on a held-out test set. We used the Tree-structured Parzen Estimator (TPE) algorithm (Watanabe, 2023) with the Optuna backend (Akiba et al., 2019) in Ray Tune (Liaw et al., 2018). The algorithm fits two Gaussian mixture models (GMMs), one to the set of parameter values associated with the best objective and another to the remaining ones. It chooses new parameters to explore by maximizing the ratio of the likelihood between these two GMMs. As such, it is a search strategy which uses results from prior tested hyperparameters to inform the next choice of hyperparameters to test. We used the TPE algorithm to search over all relevant hyperparameters above. Additionally, for MR-SDS, we used dropout for regularization with the default settings in the provided implementation. We also manually picked a good learning rate and number of epochs for training. Finally, we also made use of co-smoothing for evaluation. This holds out a set of neurons from the inference network, and computes the fit on the reconstruction of these held-out neurons.

A.4. Model Comparisons

We outline the components of MR-LFADS variants and existing communication models in Table S3. Models are compared across four criteria: (1) region-specific dynamics, (2) unsupervised inferred inputs, (3) data-constrained communication, and (4) structured information bottlenecks. Only MR-LFADS(R) incorporates all four features.

MR-LFADS(S) ablates the controller, removing inferred inputs and instead using manually specified external inputs for each region. MR-LFADS(F) and MR-LFADS(G) both communicate via latent variables not directly grounded in observed data—using factors and generator states, respectively.

MR-SDS and mp-srSLDS include region-specific dynamics but rely on external inputs and latent-variable-based messaging, without any regularization on inputs or communication. RRR infers communication from observable quantities but lacks dynamics and inputs altogether. While it enforces a rank constraint on the communication subspace, this is not equivalent to explicit regularization on messages.

Table S3. Comparison of MR-LFADS variants and existing communication models.

	Region-Specific Dynamics	Unsupervised Inferred Inputs	Data-Constrained Communication	Structured Information Bottlenecks
MR-LFADS(R)	✓	✓	✓	✓
MR-LFADS(S)	✓	×	✓	×
MR-LFADS(F)	✓	✓	×	✓
MR-LFADS(G)	✓	✓	×	✓
MR-SDS	✓	×	×	×
mp-srSLDS	✓	×	×	×
RRR	×	×	✓	×

B. Evaluation Metrics

B.1. Quantifying Effectome Similarity

In a trained MR-LFADS model, we define the inferred effectome to be a matrix of pairwise message norms, M, with element $M_{i,j}$ as the average value of $||\mu_{m,t}^{j\to i}||_2$ across all trials and timesteps (see Eq. 7). In Experiments 1-2, we compared model-inferred effectomes to the corresponding ground truth connectivity matrix M_{true}, consisting of ones and zeros to indicate the presence or absence of a communication channel in the DGN, respectively. In contrast, Experiment 3 features networks in which not all connections are actively used; in this case, we define M_{true} analogously to M, but computed using ground truth messages $m^{j\to i}$ instead of inferred messages $\mu_{m,t}^{j\to i}$. To assess similarity between M and M_{true}, we flatten these matrices into vectors—$\vec{m}$ and $\vec{m}_{\text{true}}$—and compute their cosine similarity:

$$S_{\text{cos}} = \frac{\langle \vec{m}, \vec{m}_{\text{true}} \rangle}{\|\vec{m}\|_2 \cdot \|\vec{m}_{\text{true}}\|_2} \in [0, 1], \tag{20}$$

where perfect alignment is indicated when $S_{\text{cos}} = 1$.

To visualize an inferred effectome (e.g., Fig. 2c, right), we plot arrows with line thickness indicating the corresponding scalar message norm from the inferred effectome. We thresholded and clipped these scalars before plotting to improve visual clarity. The threshold is determined using k-means clustering on the $|||\mu_{m,t}^{j\to i}||_2$ values, separating them into two groups and omitting the group with smaller values. Line thickness is capped at an upper bound for readability. All reported values elsewhere are computed without applying any thresholds, and visualizations of the unthresholded effectome are also provided in Fig. S1b, Fig. S2b.

B.2. Evaluating Information Encoded in Learned Messages

In Experiment 1, we tested whether the inferred inputs and communications encoded information about the past ground truth values, as the network's hidden units activity contains information about past inputs. A correct model should only learn the ground truth inputs or communications. For the results shown in Fig. 2d-e, right, the r-squared values were calculated with a time lag d as $R^2(\mu_{m,t}^{j\to i}, m_{t-d}^{j\to i})$.

C. Synthetic Datasets for System Identification Issues

Synthetic datasets for networks from Experiments 1-3 have 1024, 1024 and 820 total trials respectively, of which 85% is used for training, and 15% for validation. The length of each trial is 200 time steps.

C.1. Memory Network

In this synthetic network, each region A^i is modeled as an GRU network with 64 units that receives a private stimulus $s_t^i \sim \mathcal{N}(0, \mathbb{I})$ with dimensions r^i. To simulate communication channels carrying different amounts of independent information, the dimensions are set as $(r^1, r^2, r^3) = (2, 3, 4)$. Each region has a linear readout W_{out}, and the outputs are required to encode information about the history of all inputs (i.e., stimuli and communication) for up to 5 time steps, enforced using a mean squared error loss. Similarly, the messages transmitted between regions are trained to match s_{t-2}^i and are also optimized via a mean squared error loss. Additionally, each region is subjected to dynamic noise $\xi \sim \mathcal{N}(0, 0.01\,\mathbb{I})$, introduced as perturbations to the RNN activity at each time step.

After training the synthetic network, for MR-LFADS, we performed a hyperparameter sweep over the KL penalty coefficients for inferred inputs ($\beta_u \in \{0.01, 0.1, 1, 10\}$) and communication ($\beta_m \in \{0.001, 0.01\}$). The coefficient pair that resulted in the lowest held-out neuron loss, $(\beta_u, \beta_m) = (0.1, 0.01)$, was selected. Using these optimized coefficients, we ran the MR-LFADS fit across 10 different seeds, which randomizes model initialization and subsequent sampling of inferred quantities during training, but does not change the allocation of training versus validation data.

Comparing Experiments 1 and 2, it is shown that mp-srSLDS and MR-SDS activity reconstruction performance underperforms compared to the MR-LFADS variants in Experiment 1 (Fig. 2b), but not in Experiment 2 (Fig. 3b). One possible explanation for this discrepancy is the amount of information that the latent variables must encode at each time step. For example, in area A^1, the private stimulus is $s_t^1 \in \mathbb{R}^2$, and the incoming message from area A^3 is $m_t^{3\to 1} = s_{t-2}^3 \in \mathbb{R}^4$. As a result, the latent representation at time t must capture information spanning 5 time steps and 6 variables in total. Under a standard hyperparameter tuning scheme (Section A.3), latent variable models like mp-srSLDS and MR-SDS may struggle to represent all this information accurately.

C.2. Pass-Decision Network

In this synthetic network, each area is an GRU that contains 16 units. The stimulus s_t is two-dimensional and independently sampled from an exponential distribution with rate parameter of 3 time steps. Using a non-Gaussian distribution tests whether the model can learn the correct solution despite structural mismatches between the data and the model, as both the prior and posterior of inferred inputs and messages are Gaussian.

Each region has a linear readout W_{out}, whose output is required to match its corresponding latent variables (s_t for area A^P and d_t for A^D). The message sent from A^P to A^D is trained to represent s_t. Additionally, A^D must encode whether d_t is greater or less than 0 at all times, mimicking a binary decision-making process.

For the pass-decision network, a low KL penalty for inferred inputs ($\beta_u = 0.0075$) was necessary to achieve good held-out neuron loss, while the KL penalty for communication ($\beta_m = 0.001$) was set to be approximately one order of magnitude smaller. Using these coefficients, we ran the MR-LFADS fit across 10 different seeds.

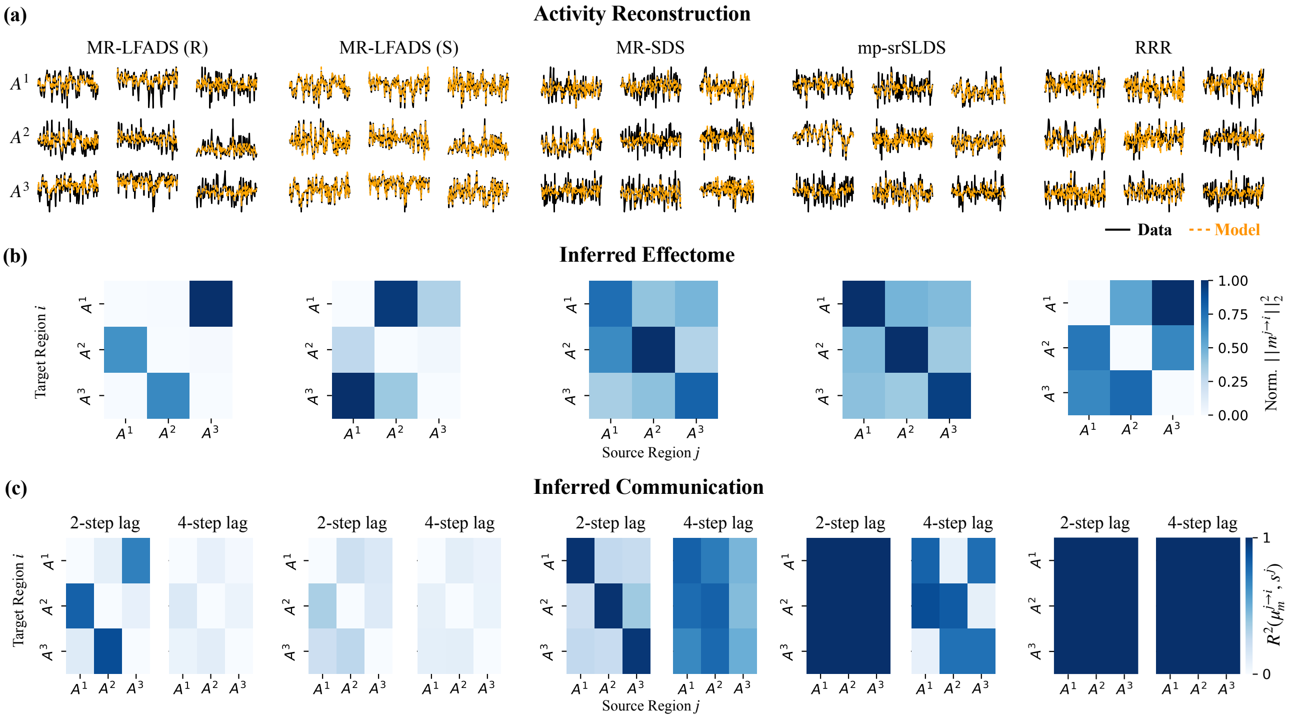

Figure S1. Memory network. (a) Example traces of model reconstruction. *Row:* example neuron from an area, *column:* different trials. (b) Inferred effectomes visualized as heatmaps. Color represents message norms normalized by the largest message norm within each model, $\max_{i,j} ||\mu_m^{j\to i}||_2$. These largest messages are, from left to right: 195, 22, 95, 1431, 5. (c). R^2 of linear prediction of ground truth messages (with 2 time step lag on the left, 4 time step lag on the right) via inferred messages. MR-LFADS(R) results (top) are replicated from Fig. 2e.

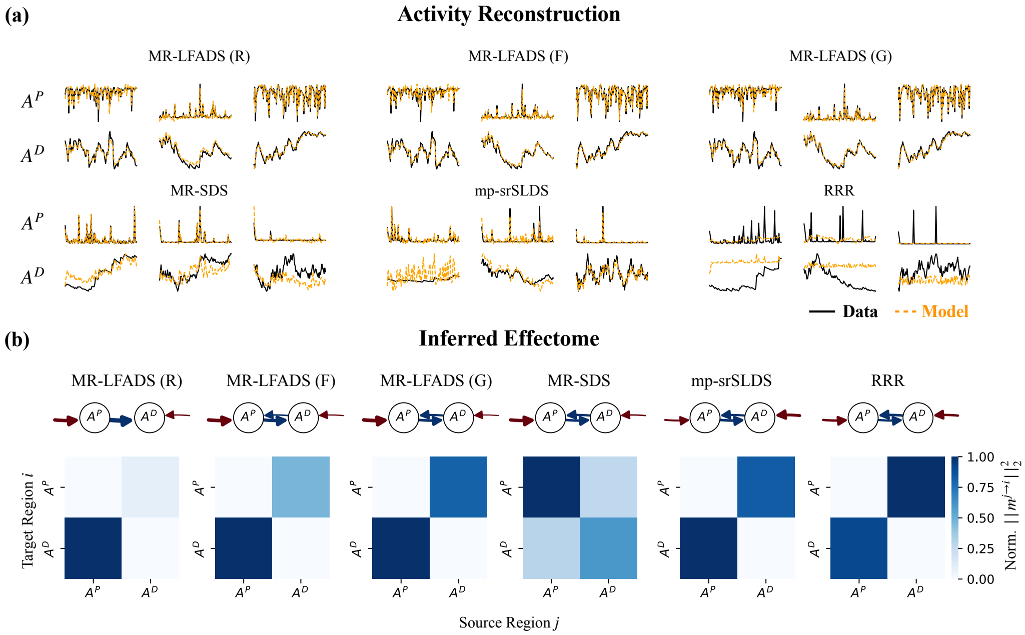

Figure S2. Pass-decision network. (a) Example traces of model reconstruction. *Row:* example neuron from an area, *column:* different trials. (b). (b) Inferred effectomes visualized as circuit diagrams and heatmaps. Color represents message norms normalized by the largest message norm within each model, $\max_{i,j} ||\mu_m^{j\to i}||_2^2$. These largest messages are, from left to right: 287, 271, 307, 143, 398, 0.6.

D. Randomly Generated Multi-Region Networks

Table S4. Task parameters for all families in multi-region data generating networks.

Task	Δ_{offset}	Δ_{delay}	t_{sacc}	θ^{resp}
Go	N/A	∞	$t_{\text{start}} + \Delta_{\text{dur}}$	$\theta^{(i)}$
Anti-Go	N/A	∞	$t_{\text{start}} + \Delta_{\text{dur}}$	$\pi + \theta^{(i)}$
Delay-Go	N/A	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{delay}}$	$\theta^{(i)}$
Delay-Anti-Go	N/A	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{delay}}$	$\pi + \theta^{(i)}$
DM1	0	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}}$	$\theta^{(1)}$
DM2	0	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}}$	$\theta^{(2)}$
MultSen DM	0	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}}$	$\theta^{(\mathrm{i})},\ i = \arg\max_i r^{(i)}$
Delay-DM1	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$	$\theta^{(1)}$
Delay-DM2	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$	$\theta^{(2)}$
Delay MultSen DM	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$	$\theta^{(\mathrm{i})},\ i = \arg\max_i r^{(i)}$
Angle	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$ if $\theta^{(1)} = \theta^{(2)}$	$\theta^{(2)}$
Anti-Angle	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$ if $\theta^{(1)} = \theta^{(2)}$	$\pi + \theta^{(2)}$
Category	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$ if $\text{sign}(\theta^{(1)}) = \text{sign}(\theta^{(2)})$	$\theta^{(2)}$
Anti-Category	[10, 20)	[30, 50)	$t_{\text{start}} + \Delta_{\text{dur}} + \Delta_{\text{offset}}$ if $\text{sign}(\theta^{(1)}) = \text{sign}(\theta^{(2)})$	$\pi + \theta^{(2)}$

We generated a distribution of networks designed to perform computational tasks inspired by Yang et al. (2019). Each network consists of either 3 or 4 regions. A connection probability $p \in \{0.5, 0.6, 0.7\}$ is specified, and the connectome is randomly drawn. To be considered valid, the connectome must meet two criteria: (1) each region must have at least one input connection and one output connection to ensure no region is redundant, and (2) all regions must be within a maximum distance of 2 steps from the output region. Once a valid connectome is generated, the network is trained on one of the computational tasks. During training, dynamic noise $\xi \sim \mathcal{N}(0, 0.01\,\mathbb{I})$ is applied to all regions, and the only loss is based on whether the output region produces the correct response. Networks that meet the performance thresholds (train accuracy > 0.8 and validation accuracy > 0.6) are selected as synthetic datasets.

For this case, since we aim to collect a distribution of results, we do not perform a hyperparameter sweep over the KL penalties. Instead, we fit one instance of each communication model to each synthetic dataset using a single random seed. The computational tasks inspired by Yang et al. (2019) are described below.

All trials are 200 time steps in length. Each task receives a fixation input $s_{\text{fix},t}$, two stimuli $s_t^{(1)} = (a^{(1)}, \theta^{(1)})$ and $s_t^{(2)} = (a^{(2)}, \theta^{(2)})$, and requires a saccade response r_t^{sacc} and an additional response $r_{\text{resp},t} = (1, \theta_{\text{resp}})$, which differ based on the specified task. Both the stimuli and responses are expressed in polar coordinates, with a resolution of 10 degrees per angle.

The tasks are described in terms of 3 different families: the go task family, context-dependent decision-making family, and matching family. For all tasks, $t_{\text{start}} \in [30, 50)$ denotes the onset of the first (and sometimes only) stimulus. The duration of all stimulus pulses in a trial is represented by $\Delta_{\text{dur}} \in [30, 50)$, while Δ_{offset} specifies the offset between the two stimuli, if applicable. The time between the last stimulus offset and the fixation cue offset is given by Δ_{delay}, where $\Delta_{\text{delay}} = \infty$ indicates that the fixation cue never disappears. The onset of the saccade is denoted as t_{sacc}.

D.1. Go Task Family

The common characteristic of tasks in this family is that only one of the stimulus channels contains the signal, which varies between trials. Depending on the specific task, the network must saccade dependently or independently of the fixation cue. The response is required to be either in the direction of the signal pulse, $\theta^{(i)}$, or in the opposite direction, $\pi + \theta^{(i)}$. The individual tasks are summarized in Table S4.

D.2. Context-Dependent Decision-Making Family

For this family of tasks, stimulus pulses occur in both channels, and the network must report either $\theta^{(1)}$ or $\theta^{(2)}$, depending on the specific task type. In some tasks, the pulses occur at different times, requiring the network to maintain memory of the stimuli. The task parameters for this family are summarized in Table S4.

D.3. Matching Family

In the matching tasks, the network determines whether to saccade based on whether the two stimulus angles "match." In the "Angle" tasks, the network saccades only if $\theta^{(1)} = \theta^{(2)}$ under the given resolution (10 degrees per angle). In the "Category" tasks, the network saccades if $\text{sign}(\theta^{(1)}) = \text{sign}(\theta^{(2)})$, meaning both angles are either positive or negative. Task details are provided in Table S4. Regardless of whether the angles match, the response is always set to report the angle $\theta^{(2)}$ (or the opposite of it).

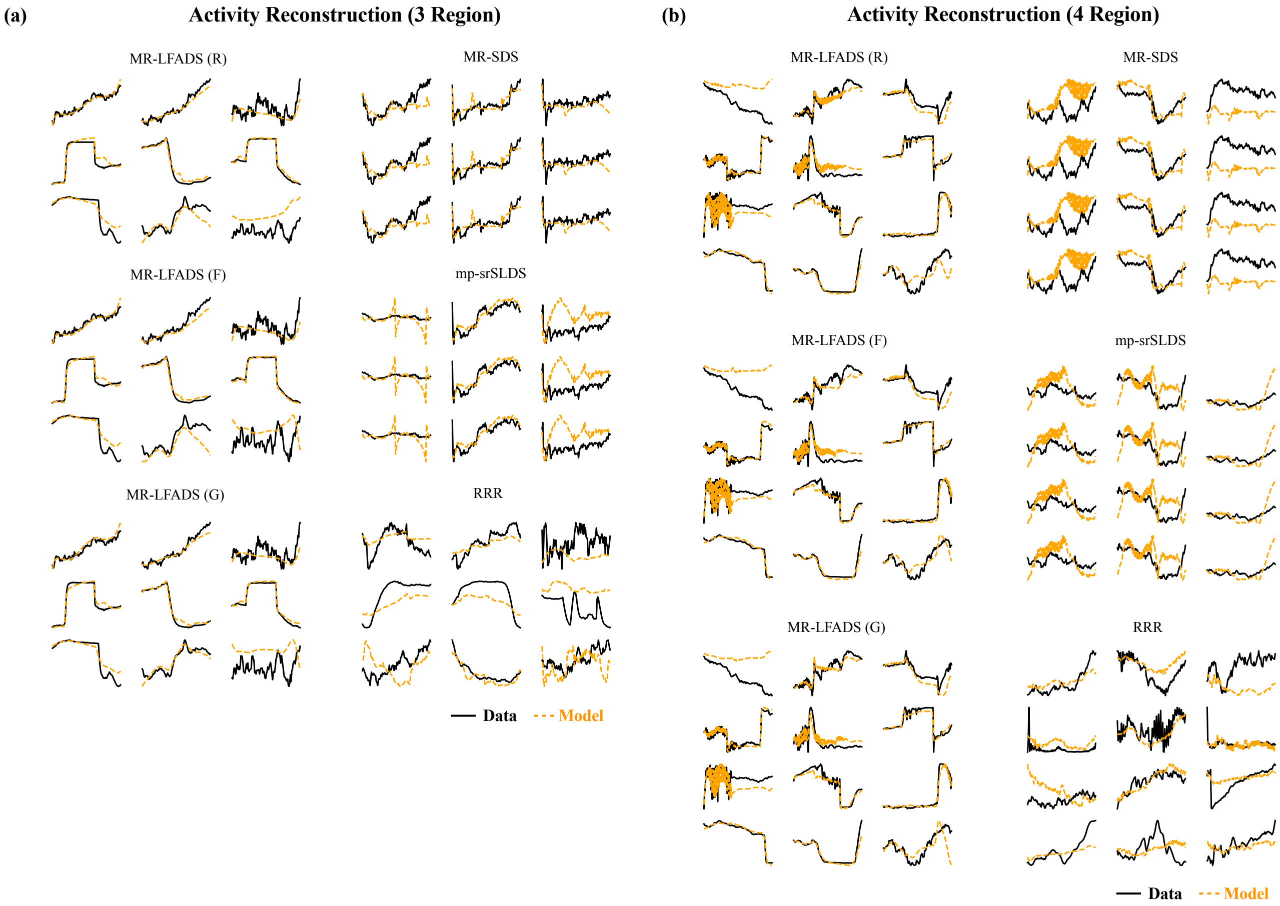

Figure S3. Randomly generated multi-region network: example traces of model reconstruction for networks with median MR-LFADS(R) performance for 3 regions (a) and 4 regions (b).

E. Implications of Constrained Architectural Choices in Restricting Message Content

E.1. Input and Message Inference with KL Penalties

KL penalties with standard Gaussian priors in variational autoencoders are known to reduce latent space dimensionality by pruning unnecessary dimensions (Dai et al., 2018; Miller et al., 2024). Consequently, with sufficiently high KL penalty coefficients (β_u, β_m), MR-LFADS is incentivized to use only the communication channels essential for data reconstruction. This effect is evident when comparing the most active channel (i.e., the one with the highest input or message norm across trials and time) to the least active ones (Fig. S4a, b). Examining input and message norms across all channels further confirms that some channels are effectively silenced (Fig. S4c).

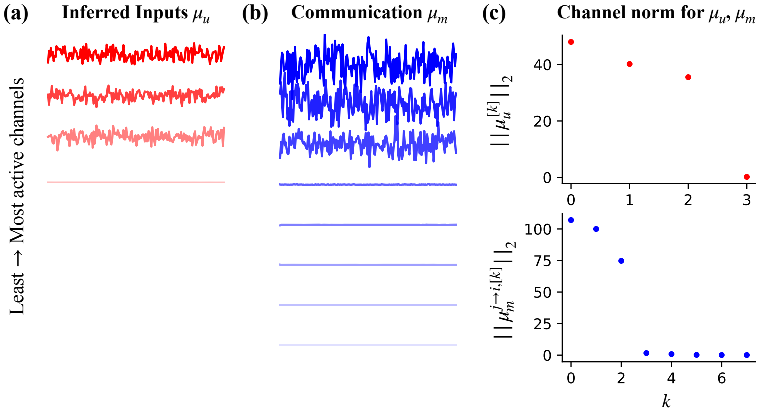

Figure S4. The effect of KL penalties on restricting information across inferred inputs and communication channels in Experiment 1. (a) Inferred input norm over time, $||\mu_{u,t}^{[k]}||_2$, for the most (red) and least (pink) active channels, where k denotes channel number. (b) Inferred message norm over time, $||\mu_{m,t}^{j\to i,[k]}||_2$, for the most (blue) and least (cyan) active channels. (c) Inferred input / message norm for area A^1.

E.2. Message Inference via Rates versus Latent Factors

Since latent factors and reconstructed rates share the same dimensionality and are related by a linear transformation, it may not be immediately clear how message inference from these variables leads to substantial differences. To investigate this, we examine MR-LFADS(F) trained on the pass-decision synthetic data. We perform SVD on projection matrices from factors to rates, W_r, and from factors to communication, W_m (Fig. S5a). Our analysis shows that W_r has small singular values, indicating that certain latent factor dimensions contribute minimally to the reconstructed rates (Fig. S5b). In contrast, W_m exhibits relatively uniform singular values (within the same order of magnitude), suggesting that all latent factor dimensions are utilized in communication (Fig. S5c). This implies that some latent factor dimensions play a role in message inference while being largely detached from rate reconstruction (Fig. S5a).

To further investigate, we re-express the factor space using the left singular vectors of W_r, denoted U. In Experiment 2, MR-LFADS(F) is shown to encode both the stimulus s and decision variable d in its $m^{P\to D}$ messages (Fig. 3e). To examine how these variables are distributed across the subspaces of U and whether this aligns with the under-constrained dimensions identified earlier, we project the latent factors f onto the subspace spanned by the top k singular vectors of U, denoted $U^{[1:k]}$, varying k from 1 to N_{msg}^D. We then compute the R^2 values for predicting s and d from the projected values f' (Fig. S5d, e, cyan lines). We repeat this process in the reverse direction, projecting onto the bottom k singular vectors instead (Fig. S5d, e, orange lines). The results reveal a clear separation: decoding accuracy for s improves more when projecting onto the top singular vectors, whereas decoding accuracy for d increases more rapidly when projecting onto the bottom singular vectors. This suggests that information not used for rate reconstruction—such as d—is preferentially encoded in the under-constrained dimensions of the latent space, reinforcing the idea that message inference utilizes latent dimensions beyond those needed for rate prediction.

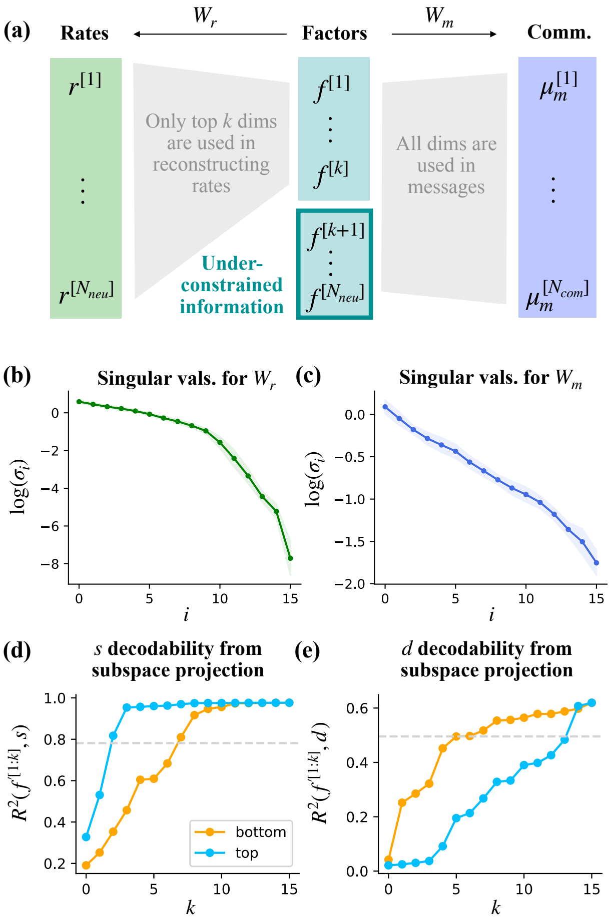

Figure S5. Potential pitfall of overexpressive models. (a) Illustration of latent factors containing unconstrained information. k is the effective rank of W_r. (b) Ranked singular values for W_r. (c) Ranked singular values for W_m. (d) R^2 for decoding s from f^P projected onto subspaces spanned by top / bottom left singular vectors of W_r of an example seed. (e) Same as (d), but for decoding d.

F. Application to Large Scale Electrophysiology Data

Data analyzed are from mouse ID:440959. For the photoinhibition experiment, we analyzed a session with the following recorded regions:

- **Anterior Lateral Motor cortex (ALM)**: MOs2/3, MOs5, and MOs6 (layer 6)
- **Thalamus (ALM, A)**: VM and VAL
- **Thalamus (Other, O)**: Anterior Ventral (AV) and Lateral Dorsal (LD)
- **Midbrain Reticular Nucleus (MRN)**: MRN
- **Superior Colliculus (SC)**: intermediate gray (SCig), intermediate white (SCiw), optic (SCop), superficial gray (SCsg), and zonal layer (SCzo)

For the model consistency experiment, we analyzed a session with the these recorded regions:

- **ALM**: secondary motor cortex, layers 2/3 and 5 (MOs2/3, MOs5)
- **Thalamus**: Ventral Medial (VM) and Ventral Anterior-Lateral (VAL)

- **MRN**: MRN

Trials were filtered to include only those with durations between 4.5 and 5.5 seconds. Each trial was binned into 500 time steps, with each bin corresponding to a 10 ms interval.

F.1. Comparison of Photoinhibition Effects

We selected one session of the data involving five brain regions previously implicated in a decision-making task, as identified in Chen et al. (2024). For both control and photoinhibition trials, we only used trials from the same condition per comparison—either left-hit or right-hit—where "hit" indicates a correct choice, and "left" or "right" refers to the correct decision side. For photoinhibition trials, we focused on those with perturbations within the delay period ($\sim$ 2 s), aligning all such trials to the photoinhibition onset. Firing rate $\bar{r}_t$ is smoothed from raw spike counts using a causal exponential filter with rate parameter of 7.1. For each region, we computed the absolute difference in trial-averaged activity between photoinhibited and control trials, averaged over neurons, to estimate the influence of photoinhibition.

F.2. Consistency of Model Inference Across Random Seeds

To ensure a fair comparison between models, we evaluated each using the same hyperparameters and same number of random seeds. We then computed all pairwise similarities between inferred effectomes and messages across seeds to assess the consistency of model inference.