INTER-REGIONAL COMMUNICATION BENCHMARK '26: SIMULATED DATASETS AND EVALUATION METRICS FOR MULTI-REGIONAL MODELS OF NEURAL POPULATION ACTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-region neural recordings have motivated a growing family of data-driven models that aim to infer the pathways and content of inter-regional communication. Evaluating these models is challenging because communication is typically unobserved, and accurate reconstruction of observed neural activity does not imply accurate recovery of unobserved communication. Model validation has therefore largely relied on synthetic datasets with ground-truth communication. However, due to the lack of standardized datasets and evaluation metrics, each existing model has been validated in a unique testbed, often designed alongside the model itself. This fragmented approach precludes systematic comparison across models and risks obscuring models' capabilities and limitations. To address this gap, we introduce a benchmark comprising 42 synthetic multi-region datasets and a suite of evaluation metrics for rigorous assessment of communication models. We demonstrate the benchmark on two communication models, revealing model limitations and conditions that impact model performance. Altogether, we believe this resource will stimulate and guide model development, and support scientific interpretation of large-scale multi-region neural recordings.

1 INTRODUCTION

A fundamental challenge in systems neuroscience is understanding how distributed brain regions coordinate their neural activity to support flexible behavior (Breakspear, 2017; Steinmetz et al., 2019; Vyas et al., 2020). Advances in large-scale neural recording technologies, such as high-density electrophysiology (Trautmann et al., 2023) and wide-field calcium imaging (Nietz et al., 2022), now enable simultaneous observation of neural population activity across multiple brain regions. These recordings have motivated the development, application, and interpretation of *communication models*, which seek to infer how neural populations interact across regions, including which regions interact and what signals are exchanged (Semedo et al., 2019; Glaser et al., 2020; Karniol-Tambour et al., 2024; Liu et al., 2025). Evaluating these models is challenging because the inter-regional communication they seek to recover is typically not directly observed in experimental recordings. Consequently, models are often validated on synthetic datasets that provide ground-truth access to communication. However, there exists no standardized set of datasets and evaluation metrics for validating communication models. As a result, each existing communication model has largely been validated in its own unique testbed, with synthetic datasets and evaluation criteria designed alongside each model. Their performance therefore cannot be systematically compared against other models, and models may also be prone to distinct failure modes that may go undetected if the testbed does not span the conditions that expose them.

A rigorous evaluation of communication models should test whether models infer the correct *causal interactions*. Typical multi-region recordings reflect multiple unobservable influences, including local-region dynamics, communication from recorded regions, and inputs from unrecorded regions. These influences are difficult to disentangle from observational neural recordings alone. Although perturbation experiments remain the gold standard for identifying causal influences (Kang & Druckmann, 2020; Nejatbakhsh & Wang, 2025), models that more reliably recover causal structure from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

observational data could substantially narrow the space of candidate interactions, informing which interactions to test with targeted perturbation experiments. However, existing synthetic datasets may not adequately test this capability, as they typically use simplified data-generating systems that may not fully capture the challenges of experimental recordings. These simplifications include restricting connectivity to feedforward interactions, or using canonical low-dimensional systems that do not meaningfully relate to neural computation, therefore providing a limited proxy for goal-directed multi-region neural population dynamics.

Motivated by these limitations, we introduce the **Inter-Regional Communication Benchmark 2026 (IRCB '26)**, a standardized benchmark comprising 42 synthetic multi-region datasets and a suite of metrics for evaluating communication models (Figure 1). Each dataset is produced by a data-generating network (DGN), a set of interconnected recurrent neural networks (RNNs), where each RNN represents a brain region and its hidden units represent individual neurons. Each DGN is trained to perform one of 22 cognitive-neuroscience tasks (Yang et al., 2019), producing noisy, recurrent, distributed, and goal-directed computations that capture important properties of interacting neural populations. We additionally provide partial-observability conditions in which subsets of neurons or entire regions are withheld from observations, reflecting constraints encountered in experimental recordings (Iwasaki et al., 2018; Qian et al., 2024).

To comprehensively assess recovery of the ground-truth systems underlying the DGNs, we provide metrics that evaluate different aspects of the inferred systems. It is often instructive to assess a model's ability to reconstruct data, so we first assess neural activity reconstruction by measuring how closely a model's reconstructed rates match the ground-truth rates of the synthetic system. We then assess whether the model recovers the correct causal flow of effects across its inter-regional connectivity, henceforth termed the *effectome* (Pospisil et al., 2024). Critically, the effectome may be dynamic with state- and task-dependence (Chen et al., 2024), and may therefore differ from the anatomical connectome. Finally, we evaluate whether the inferred communication captures the ground-truth communication content without encoding additional information absent from ground truth. We demonstrate the benchmark on two communication models to illustrate the analyses and model comparisons enabled by our framework. IRCB '26 therefore establishes a standardized testbed for comparing communication models under rigorous evaluation criteria, and can help guide the development of methods that more reliably recover inter-regional interactions from the growing volume of large-scale neural recordings.

Code for dataset generation and model evaluation is available at `https://github.com/team-ircb/IRCB-26`, and the released datasets are available at `https://huggingface.co/datasets/ircb/IRCB-26`.

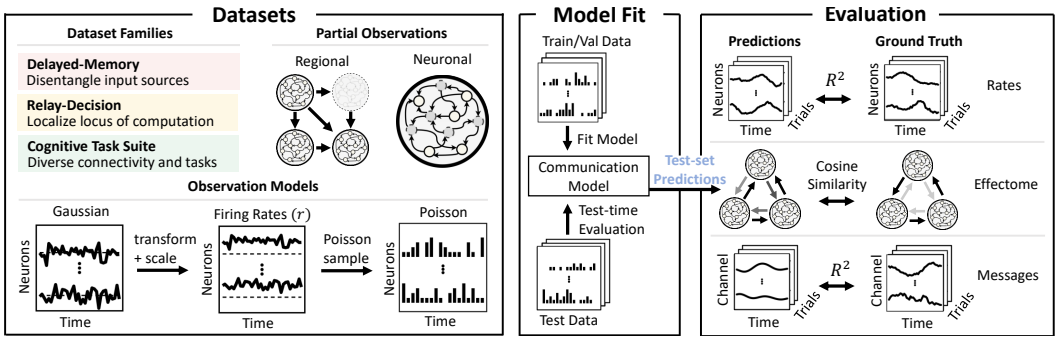


Figure 1: **Overview of IRCB '26.** ***Datasets.*** Synthetic neural activity is generated from one of three dataset families, DELAYED-MEMORY, RELAY-DECISION, and the COGNITIVE TASK SUITE, with Gaussian or Poisson observations and full or partial observability. ***Model fit.*** Models are fit on training/validation data and evaluated on held-out test data. ***Evaluation metrics.*** Predicted neural activity, effectomes, and inter-regional messages are compared with ground truth.

2 Related Work

Benchmarks for neural population modeling. While no standardized benchmarks have been established for evaluating inter-regional communication models, several benchmarks have been developed for evaluating single-region models of neural population dynamics. The Neural Latents Benchmark (NLB '21) provides neural recordings spanning motor, sensory, and cognitive tasks, and evaluates a model's ability to predict the activity of held-out neurons and behavioral variables (Pei et al., 2021). However, because the underlying dynamics are not directly accessible, these metrics cannot directly evaluate their recovery. Although some NLB datasets do include recordings from multiple cortical areas, the benchmark does not evaluate recovery of their interactions, as neural recordings do not provide ground truth for the signals communicated between regions.

To enable direct evaluation of the underlying dynamical system, the Computation-through-Dynamics Toolkit (CtDToolkit) uses synthetic datasets generated by task-trained dynamical systems, which provides access to the ground-truth data-generating system (Versteeg et al., 2026). This enables direct evaluation of inferred dynamics through latent trajectories, fixed points, and task-relevant representations. However, while it shares our goal of distinguishing accurate neural activity reconstruction from faithful recovery of the underlying system, it focuses on recovering the dynamics of a single system rather than the interactions between interconnected dynamical systems.

Multi-region communication models. Existing multi-region methods seek to characterize interactions between neural populations using a range of statistical and dynamical approaches. Static methods such as reduced-rank regression (RRR) identify low-dimensional subspaces through which one region's activity predicts another's (Semedo et al., 2019; MacDowell et al., 2025). Temporal latent-variable approaches, including Delayed Latents Across Groups (DLAG) (Gokcen et al., 2022) and its multi-population extension mDLAG (Gokcen et al., 2023), separate within- and across-region latent activity and use temporal delays to characterize directed signal flow. Another class of approaches explicitly models each region as a dynamical system and couples these systems through inter-regional interactions, including multi-population sticky recurrent Switching Linear Dynamical System (mp-srSLDS) (Glaser et al., 2020), Multi-Region Switching Dynamical Systems (MR-SDS) (Karniol-Tambour et al., 2024), and Multi-Region Latent Factor Analysis via Dynamical Systems (MR-LFADS) (Liu et al., 2025). However, despite these methodological advances, multi-region communication models are still largely validated in separate, method-specific testbeds, limiting systematic comparison and rigorous evaluation under shared conditions.

3 Datasets

3.1 Overview

Our benchmark includes three classes of DGNs that are constructed to evaluate a communication model's ability to disentangle the effects of local-region dynamics, communication between recorded regions, and inputs from unrecorded regions (Figure 2). The first two, the Delayed-Memory and Relay-Decision networks, are designed to impose targeted challenges. Delayed-Memory tests whether models correctly distinguish inputs from recorded and unrecorded regions, while Relay-Decision tests whether models correctly attribute computations to the regions in which they occur. In both networks, the inter-regional communication is explicitly prescribed, providing controlled ground truth for evaluating recovery of its content.

The third class of DGNs, the Cognitive Task Suite, consists of 40 networks trained on 20 cognitive tasks adapted from Yang et al. (2019), with two networks per task and an independently sampled, fixed anatomical connectivity structure for each network. This connectivity determines which inter-regional communication pathways are available, while the messages transmitted along those pathways emerge freely through task training. This design enables assessment of communication models across diverse connectivity structures and emergent goal-directed dynamics.

Across all three classes, each DGN consists of interconnected vanilla RNN modules, each representing an abstract brain region. The hidden state $\mathbf{h}_t^i \in \mathbb{R}^{n^i}$ represents the activity of n^i units in region i at

time t and updates according to

$$\begin{aligned}\mathbf{h}_t^i &= \tanh\left(W_{\mathrm{hh}}^i \mathbf{h}_{t-1}^i + W_{\mathrm{sh}}^i \mathbf{s}_t^i + \sum_{j\in\mathcal{S}^i} W_{\mathrm{mh}}^{j\to i}\mathbf{m}_t^{j\to i} + \mathbf{b}^i\right) + \boldsymbol{\epsilon}_t^i,\\ \mathbf{m}_t^{j\to i} &= W_{\mathrm{hm}}^{j\to i}\mathbf{h}_t^j + \boldsymbol{\xi}_t^j,\end{aligned} \tag{1}$$

where $\mathbf{s}_t^i \in \mathbb{R}^{d_s}$ denotes the external input to region i at time t, $\mathbf{m}_t^{j\to i} \in \mathbb{R}^{d_m}$ denotes the message from region j to region i at time t, and $\mathcal{S}^i$ denotes the set of source regions projecting to region i. We additionally constrain d_m to be low-dimensional, consistent with evidence that inter-regional interactions are typically mediated through low-dimensional population subspaces (Semedo et al., 2019; MacDowell et al., 2025). The matrices W_{hh}^i, W_{sh}^i, and $W_{\mathrm{mh}}^{j\to i}$ map hidden-unit activity, external inputs, and incoming messages, respectively, into the hidden-state update. The matrix $W_{\mathrm{hm}}^{j\to i}$ projects the high-dimensional activity of region j into the low-dimensional communication signal sent to region i. $\boldsymbol{\epsilon}_t^i$ and $\boldsymbol{\xi}_t^j$ denote IID Gaussian dynamical noise and communication noise, added after the nonlinear regional update and to the communication signal, respectively. Additional details on DGN training, including optimization and hyperparameters, are provided in Appendix A.

To obtain Poisson spike-count observations with realistic signal-dependent noise, we first standardize each region's hidden-state activity to zero mean and unit variance, accounting for differences in activity scale across regions. The standardized activity $\tilde{\mathbf{h}}$ is then translated and scaled to obtain the firing rate

$$\mathbf{r} = r_{\mathrm{scale}}\frac{\tilde{\mathbf{h}} + 1}{2},$$

where $r_{\mathrm{scale}} = 40\,\mathrm{Hz}$. For each $\Delta t = 10\,\mathrm{ms}$ time bin, we then sample

$$y_t \sim \mathrm{Poisson}(\max(0, \mathbf{r}_t \Delta t)).$$

This transformation preserves the underlying DGN dynamics while introducing stochastic, signal-dependent variability in the observed spike counts. For additional details, see Appendix B.

We further include partial-observation variants that withhold subsets of neurons or entire regions from communication models to reflect experimental limitations. Because the observation constraints are applied after DGN training, the effects of unobserved neurons and regions are retained in the observed activity. Each dataset includes four neuronal observation masks corresponding to 25%, 50%, 75%, and 100% neuronal coverage, where 100% represents full observability. For datasets in the COGNITIVE TASK SUITE, we additionally provide two regional observation masks that withhold the activity of an entire region (Appendix C).

To support standardized evaluation across models, all datasets and observation variants are released in a common `HDF5` format with fixed training, validation, and test sets. Communication models should be trained using only the observed neural activity, while ground-truth inter-regional connectivity and communication signals should be reserved for evaluation. This separation ensures that performance reflects communication recovery from neural activity, rather than access to ground truth. Models are evaluated on the held-out test set by submitting reconstructed neural activity and inferred communication outputs, with additional quantities such as external inputs evaluated when available. Further details on datasets, attributes, and submission formats are provided in Appendix E.

3.2 DELAYED-MEMORY TASK

Many computations require task-relevant information to be maintained over short timescales and coordinated across distributed brain regions (Funahashi, 2017; Rezayat et al., 2021; Clarke-Williams et al., 2024). To simulate these computations, we designed a 3-region network where each region must remember time-varying inputs and communicate aspects of this memory to a downstream region (Figure 2a). Each region A^i receives a private time-varying stimulus $s_t^i \sim \mathcal{N}(0, \mathbb{I})$ and communicates a 2-time-step delayed signal s_{t-2}^i to its downstream region. Each region is also trained to maintain a representation of the preceding 2 time steps of all received inputs, including its private stimulus and incoming communication signals. Consequently, regional activity combines both newly received information and locally maintained history, while outgoing communication messages carry a prescribed delayed component.

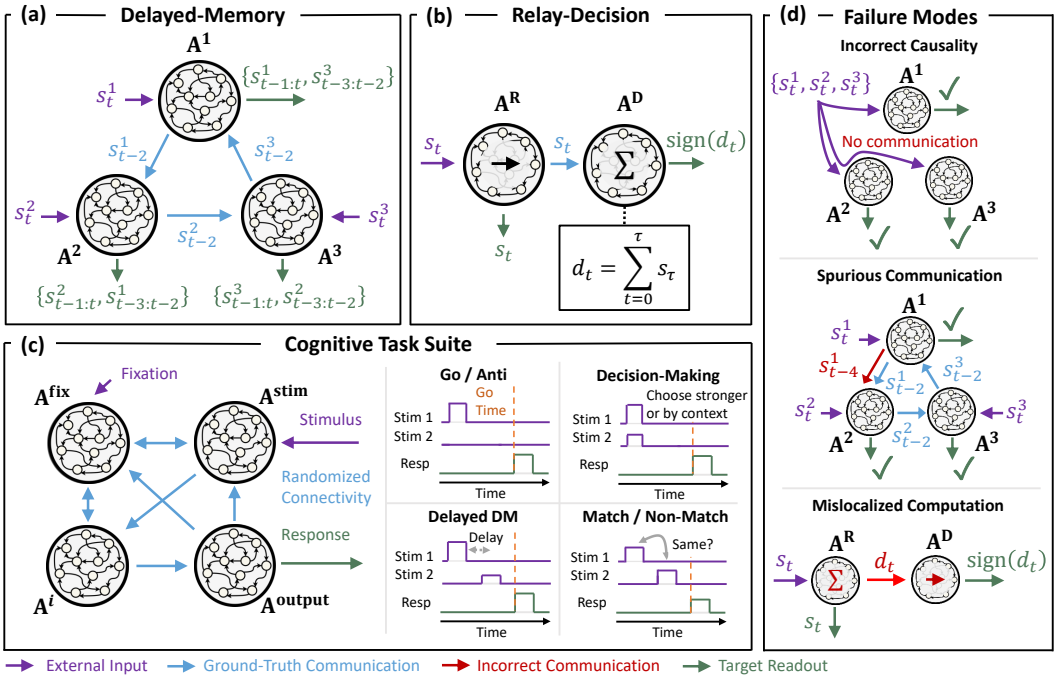


Figure 2: **IRCB '26 task architectures and failure modes. (a, b, c–*left*)** Ground-truth network architectures and task-specific communication requirements for the three DGN classes, DELAYED-MEMORY, RELAY-DECISION, and the COGNITIVE TASK SUITE. **(c–*right*)** Example trial structures for the COGNITIVE TASK SUITE task families. **(d)** Potential failure modes of communication models while recovering inter-regional communication.

This specific network construction exposes two potential failure modes for communication models. First, models that do not account for inputs from unrecorded regions and instead provide experimental stimuli directly to all modeled regions may accurately reconstruct neural activity without using inter-regional communication, leading to incorrect causal attribution (Figure 2d, top). Second, because correlated information is broadly distributed across regions, models may mistake shared representations for inter-regional interactions and infer spurious communication pathways (Figure 2d, middle). Successful modeling of DELAYED-MEMORY therefore requires correctly identifying what information originates from each recorded or unrecorded region, and distinguishing information that is communicated between regions from information that is maintained locally.

3.3 RELAY-DECISION TASK

Localizing decision-making computations across brain regions, and determining how different regions contribute to, transform, and relay decision-related signals, has been a longstanding question in systems neuroscience (Gold & Shadlen, 2007; Bollimunta & Ditterich, 2012; Hanks et al., 2015; Brody & Hanks, 2016b). To test a communication model's ability to identify the locus of computation, we designed a 2-region network where one region passively relays information while the other region integrates evidence to form a decision (Figure 2b). The relay region, A^R, receives a time-varying stimulus s_t and relays it to the decision region A^D through message $m_t^{R\to D} = s_t$. The decision region then integrates the received signal into a decision variable $d_t = \sum_{\tau=0}^{t} s_\tau$ and computes a choice as the sign of the accumulated decision variable. The message carries instantaneous evidence, while the downstream decision-region dynamics accumulates the evidence and transforms it into a choice. This separation provides ground truth for evaluating whether models recover both the transmitted signal and the locus of decision computation. Importantly, a model may incorrectly attribute the integration computation to A^R, infer the integrated signal as the communication content, and characterize A^D as merely representing rather than computing this signal (Figure 2d, bottom). Because communication and local-region dynamics are unobserved, their contributions may be

misattributed during model fitting, such that observed neural activity is accurately reconstructed via incorrect circuit-level mechanisms. Successful modeling of RELAY-DECISION therefore requires identifying the pathway from A^R to A^D, recovering the content transmitted along this pathway, and correctly localizing relay and decision processes to the correct regions.

3.4 COGNITIVE TASK SUITE

DELAYED-MEMORY and RELAY-DECISION predefine the computational role of each region and the content of inter-regional communication. In the brain, however, local-region dynamics and inter-regional communication do not necessarily have such modular roles, but may instead jointly organize in a flexible and distributed manner to support goal-directed behavior (Vyas et al., 2020; Javadzadeh & Hofer, 2022; Wang, 2022). Moreover, different networks in the brain exhibit diverse architectures and perform different computations, including routing sensory inputs, maintaining information, comparing stimuli, and forming decisions (Brody & Hanks, 2016a; Christophel et al., 2017; Takeda et al., 2018). Together, these considerations motivate a broader testbed where communication models are evaluated across a diverse range of networks with different architectures and task demands.

To capture this broader regime, we trained 40 four-region networks, each performing one of 20 cognitive neuroscience-inspired tasks adapted from Yang et al. (2019), with two networks per task, each having an independently sampled connectivity structure. These tasks include reaction-time and delayed-response tasks, context-dependent and delayed decision-making, and match/non-match computations (Figure 2c, right). For all networks, region A^{fix} receives a fixation signal, while A^{stim} receives two task-specific stimulus channels represented in polar coordinates (Figure 2c, left). The output region A^{out} then produces a task-dependent response, comprising a response direction and a saccade signal. Additional task details are provided in Appendix F.

4 EVALUATION METRICS

Neural activity reconstruction. Although accurate neural activity reconstruction does not imply faithful recovery of the underlying system, it is often instructive to understand a model's ability to reconstruct the data as well as generalize beyond it. We therefore report reconstruction performance separately for held-in and held-out neurons, where held-in neurons are provided to the communication model and held-out neurons are unobserved during fitting, requiring the model to infer their activity. We quantify rate reconstruction quality using standard R^2 between the true underlying rates $\mathbf{r}$ and the reconstructed rates $\hat{\mathbf{r}}$, denoted $R^2(\hat{\mathbf{r}}, \mathbf{r})$. We additionally report McFadden's pseudo-R^2 for the observed spike counts relative to Poisson distributions specified by the reconstructed rates $\hat{\mathbf{r}}$. Because this measure depends on a Poisson observation assumption and may therefore be sensitive to differences in models' observation distributions, we treat it as a supplementary metric.

Effectome similarity. We evaluate whether models recover the correct causal, directional communication pathways between observed regions. For the ground-truth DGN, we quantify the strength of communication from region j to region i as $\left(\sum_{k=1}^{K}\sum_{t=1}^{T}||\mathbf{m}_{t,k}^{j\to i}||_2^2\right)^{1/2}$, where $\mathbf{m}_{t,k}^{j\to i} \in \mathbb{R}^{d^m}$ denotes the communication signal from region j to region i at time $t = 1, \ldots, T$ on trial $k = 1, \ldots, K$, across d^m communication channels. We use these values to construct the effectome matrix M, where $M_{ij} = ||\mathbf{m}^{j\to i}||_2$. We analogously construct the inferred effectome, $\hat{M}$, from the communication model's inferred communication, $\hat{\mathbf{m}}$. Finally, we quantify effectome recovery using the cosine similarity between the vectorized inferred and ground-truth effectomes, $\text{vec}(\hat{M})$ and $\text{vec}(M)$. A score of 1 indicates perfect alignment, while lower scores reflect missing, spurious, or incorrectly weighted communication pathways.

For models that provide an explicit mapping from communication to the inputs of the target region's dynamics, we additionally compute a dynamically weighted effectome. In our ground-truth RNNs, communication from region j to region i enters the target region through input weights $W_{\text{mh}}^{j\to i}$ (see Eq 1). We therefore quantify its contribution to the target dynamics as $\left(\sum_{k=1}^{K}\sum_{t=1}^{T}\|W_{\text{mh}}^{j\to i}\mathbf{m}_{t,k}^{j\to i}\|_2^2\right)^{1/2}$. This metric more directly reflects the influence of communicated signals on the target region's dynamics, but is only applicable to models that explicitly parameterize how communication enters those dynamics.

Single-trial communication recovery. The inferred communication should capture the ground-truth time-varying communication content. For each pathway $j \to i$, we concatenate the ground-truth communication vectors $\mathbf{m}_{t,k}^{j \to i}$ across trials and time to form $X^{\mathbf{m}} \in \mathbb{R}^{KT \times d^m}$, and similarly concatenate the inferred communication to form $X^{\hat{\mathbf{m}}} \in \mathbb{R}^{KT \times d^{\hat{m}}}$, where $d^{\hat{m}}$ is the dimensionality of the inferred communication. We then linearly decode the ground-truth communication $X^{\mathbf{m}}$ from the inferred communication $X^{\hat{\mathbf{m}}}$. We define the R^2 score of the linear regression as the model's *communication fidelity*, denoted $R^2(\hat{\mathbf{m}}, \mathbf{m})$.

For models that explicitly infer inputs from unobserved regions, we can quantify *input fidelity* as the R^2 score of the inferred inputs $\hat{\mathbf{s}}$ predictions relative to ground-truth inputs $\mathbf{s}$, denoted $R^2(\hat{\mathbf{s}}, \mathbf{s})$.

Single-trial communication specificity. Inferred communication may have high fidelity while also containing additional information not present in the ground-truth communication. To quantify this additional information, we linearly decode the inferred communication $X^{\hat{\mathbf{m}}}$ from the ground-truth communication $X^{\mathbf{m}}$. We define a model's *communication specificity* as the R^2 score of the linear regression, $R^2(\mathbf{m}, \hat{\mathbf{m}})$, where a low value indicates that a substantial fraction of the inferred communication cannot be explained by the true communication. Similarly, for models that explicitly infer external inputs, we quantify *input specificity* as $R^2(\mathbf{s}, \hat{\mathbf{s}})$.

For DELAYED-MEMORY and RELAY-DECISION, the predefined communication content allows us to directly test whether models recover the intended content without encoding additional information. In DELAYED-MEMORY, stimulus history is represented broadly across the network, making information from multiple stimulus lags readily available to the communication model. We therefore test whether the inferred communication contains only the ground-truth content, s_{t-2}^i, or additionally encodes information from other stimulus lags by computing $R^2(\hat{\mathbf{m}}_t, \mathbf{m}_{t-\tau})$ for $\tau \neq 2$. In RELAY-DECISION, mislocalizing the decision computation to the relay region would cause the inferred communication to encode the decision variable d rather than the ground-truth stimulus s. We therefore directly test for this failure mode by decoding d from the inferred communication and computing $R^2(\hat{\mathbf{m}}, d)$.

5 RESULTS

Using IRCB '26, we benchmark two communication models: reduced-rank regression (RRR), a linear method for identifying low-dimensional predictive subspaces between neural populations (Semedo et al., 2019; MacDowell et al., 2025), and Multi-Region Latent Factor Analysis via Dynamical Systems (MR-LFADS), a nonlinear latent dynamical systems model that jointly infers region-specific dynamics, inter-regional communication, and other unobserved inputs (Liu et al., 2025). Table 1 summarizes model performance across neural activity reconstruction, effectome similarity, as well as communication and input fidelity and specificity. Model training details are provided in Appendix G, and per-task COGNITIVE TASK SUITE results are reported in Appendix H.

Neural activity reconstruction. While the reconstructed neural activity from both models correlates with the true rates, MR-LFADS achieves significantly better performance than RRR across all datasets for both held-in and held-out neurons. Notably, within the COGNITIVE TASK SUITE, reconstruction performance varies substantially across DGNs for RRR, indicating that neural activity recovery difficulty depends on each DGN's connectivity, task-specific dynamics, and communication structure.

Effectome similarity. MR-LFADS also outperforms RRR in terms of effectome similarity, particularly in the DELAYED-MEMORY and RELAY-DECISION networks. In the DELAYED-MEMORY network, stimulus information is widely available across regions, resulting in highly correlated neural activity that leads RRR to infer communication in both directions and introduce spurious pathways. Conversely, in the RELAY-DECISION network, the communicated stimulus s is only weakly correlated with the downstream decision variable d produced by its integration. Despite the causal influence of s on d, RRR infers little to no connectivity between the regions. In contrast, MR-LFADS accurately recovers the ground-truth effectome across both datasets. However, its effectome similarity is lower on the COGNITIVE TASK SUITE than on the other two DGN classes, potentially reflecting the more complex connectivity structures of the networks in the suite.

Table 1: **Example communication models evaluated on IRCB '26**. Rate and spike reconstruction are reported separately for held-in and held-out neurons and averaged across regions. Communication and input recovery are reported as (fidelity, specificity). Input recovery is omitted for RRR, as it does not infer external inputs. COGNITIVE TASK SUITE values are reported as mean ± standard deviation across 20 single-task networks. 50% denotes observing a randomly selected half of the neurons in each region, while A^{fix}-out denotes withholding all neural activity from region A^{fix}.

Dataset	Model	Rate Recon.		Spike Recon.		Effectome		Signal Recovery	
		Held-in	Held-out	Held-in	Held-out	Norm	Dynamic	Communication	Input
DELAYED MEMORY	RRR	0.254	0.242	-0.051	-0.027	0.689	0.619	(**0.724**, 0.123)	–
	MR-LFADS	**0.634**	**0.646**	**0.054**	**0.078**	**0.968**	**0.980**	(0.704, **0.911**)	(0.934, 0.946)
DELAYED MEMORY (50%)	RRR	0.209	0.163	-0.028	-0.040	0.698	0.693	(0.573, 0.094)	–
	MR-LFADS	**0.479**	**0.575**	**0.100**	**0.029**	**0.766**	**0.831**	(**0.763**, **0.885**)	(0.774, 0.141)
RELAY DECISION	RRR	0.412	0.381	-0.013	0.024	0.492	0.153	(**0.906**, 0.766)	–
	MR-LFADS	**0.913**	**0.902**	**0.119**	**0.101**	**0.914**	**0.801**	(0.811, **0.850**)	(0.939, 0.411)
RELAY DECISION (50%)	RRR	0.345	0.312	0.015	0.020	0.519	0.190	(**0.839**, 0.734)	–
	MR-LFADS	**0.838**	**0.852**	**0.110**	**0.096**	**0.853**	**0.853**	(0.662, **0.845**)	(0.858, 0.467)
COGNITIVE TASK SUITE	RRR	-0.235 ± 0.616	0.170 ± 0.373	-0.073 ± 0.033	-0.014 ± 0.018	0.788 ± 0.044	0.731 ± 0.098	(0.754, 0.690) ± (0.070, 0.044)	–
	MR-LFADS	**0.585** ± 0.152	**0.691** ± 0.171	**0.280** ± 0.118	**0.118** ± 0.024	**0.817** ± 0.053	**0.770** ± 0.086	(**0.971**, **0.962**) ± (0.017, 0.025)	(0.704, 0.204) ± (0.123, 0.307)
RT Go	RRR	-0.360	0.067	-0.100	-0.034	0.815	0.802	(0.877, 0.721)	–
	MR-LFADS	**0.809**	**0.799**	**0.146**	**0.147**	**0.832**	**0.853**	(**0.994**, **0.977**)	(0.366, 0.007)
RT Go (50%)	RRR	-0.110	0.277	-0.076	0.006	0.794	0.770	(0.785, 0.622)	–
	MR-LFADS	**0.714**	**0.797**	**0.535**	**0.132**	**0.807**	**0.822**	(**0.978**, **0.895**)	(0.308, 0.005)
Ctxt DM 1	RRR	0.141	0.301	-0.040	-0.013	0.763	0.748	(0.721, 0.724)	–
	MR-LFADS	**0.541**	**0.829**	**0.272**	**0.118**	**0.816**	**0.750**	(**0.976**, **0.980**)	(0.706, 0.010)
Ctxt DM 1 (A^{fix}-out)	RRR	0.449	0.423	0.008	0.017	0.356	0.390	(0.737, 0.715)	–
	MR-LFADS	**0.483**	**0.826**	**0.354**	**0.101**	**0.454**	**0.498**	(**0.951**, **0.981**)	(0.763, 0.013)

Communication fidelity and specificity. RRR recovers some ground-truth messages with high fidelity, particularly in the RELAY-DECISION network, despite inferring weak communication between regions and consequently achieving low effectome similarity. This discrepancy arises because fidelity measures recovery of communication content rather than magnitude, in which even a weak inferred signal can preserve the stimulus information carried by the ground-truth message.

Additionally, despite the high fidelity, RRR's inferred communication contains substantial information not present in the ground-truth messages, resulting in low communication specificity. In contrast, MR-LFADS achieves both high communication fidelity and specificity across all DGNs. Notably, however, MR-LFADS exhibits low input fidelity across the COGNITIVE TASK SUITE. This may reflect a mismatch between the ground-truth input structure and the prior imposed on the inferred inputs in MR-LFADS, highlighting an open question in LFADS-type models regarding how the choice of input prior shapes inferred inputs (Sussillo et al., 2016; Liu et al., 2025). Together, these results highlight the value of evaluating multiple aspects of recovery, revealing model limitations not apparent from any single metric.

Partial Observations. When only 50% of neurons are observed, both models generally show only modest degradation in neural activity reconstruction and communication recovery, although the extent of degradation varies across datasets and metrics. For instance, MR-LFADS's effectome similarity remains similar for the RELAY-DECISION and RT GO tasks, but drops significantly for the

DELAYED-MEMORY and CTXT DM 1 tasks. Communication fidelity decreases substantially for RRR in DELAYED-MEMORY, whereas MR-LFADS is more strongly affected in RELAY-DECISION. These results show that partial observability can have different effects across connectivity structures, task types, and even communication models, further highlighting the importance of evaluating models across a wide range of datasets.

6 DISCUSSION

Communication models are increasingly used to estimate and interpret interactions across brain regions, yet their evaluations remain fragmented across method-specific datasets and metrics. IRCB '26 addresses this gap through two key components. First, it provides multi-region datasets with ground-truth communication spanning diverse connectivity structures and computations, creating challenging conditions for communication models. Our results reveal distinct model limitations across these datasets, with model performance varying across connectivity structures and task dynamics. Second, we provide metrics that separately evaluate neural activity reconstruction and recovery of communication pathways and content. This separation is critical because success in one does not imply success in the others. Models can accurately reconstruct neural activity while failing to recover the underlying communication, or recover communicated content while misidentifying the effectome.

An important limitation of the current benchmark is that our metrics do not directly evaluate recovery of region-specific dynamics. Such recovery can be examined through latent trajectories and fixed-point structure, but defining a general quantitative comparison is challenging for high-dimensional nonlinear systems whose inferred state representations may differ substantially from ground truth. Recent approaches such as Dynamical Similarity Analysis (DSA) provide a promising framework for comparing dynamics across such representations (Ostrow et al., 2023). Future extensions of IRCB '26 could incorporate such metrics alongside fixed-point and trajectory analyses to more directly evaluate dynamical recovery (Golub & Sussillo, 2018).

Furthermore, we can broaden the scope of IRCB '26 to capture additional forms of neural computation, such as expanding the COGNITIVE TASK SUITE beyond the cognitive tasks considered here to include motor-control networks coupled to biomechanical effectors (Codol et al., 2024), introducing closed-loop sensorimotor feedback and biomechanical dynamics that are absent from the current DGNs. Finally, the benchmark could be extended to additional experimental recording conditions, including datasets with targeted perturbations. This would enable inferred interactions to be compared with measured intervention effects, providing a more direct test of whether inferred communication reflects causal influence. The benchmark could also incorporate session-stitching settings, in which interacting regions are recorded across separate sessions rather than simultaneously, reflecting an important practical constraint of large-scale neural recordings.

Altogether, IRCB '26 aims to advance communication modeling toward more reliable interpretations of large-scale multi-region recordings. By identifying where models succeed and fail to recover underlying interactions, the benchmark can inform the development of more reliable methods, while also helping generate targeted hypotheses for causal testing through perturbation experiments.

AI USE STATEMENT

In this work, we used generative AI tools to assist with generating synthetic datasets, cleaning and reformatting datasets, and verifying implementation of evaluation metrics. We have not used generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, or interpret results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs were not applicable to this work. Additionally, we used generative AI tools to summarize and analyze existing literature, source and search for information, refine wording and sentence structure, and format references. We have reviewed all AI-assisted work, including by verifying LLM-generated claims against the cited literature, checking AI-assisted code, and inspecting reformatted datasets. We take responsibility for the final content of this work, including all text, claims, code, datasets, and other artifacts produced with the aid of generative AI.

Reproducibility statement

To facilitate reproducibility, we release the IRCB '26 datasets, code for dataset generation and benchmark evaluation, and all hyperparameter and training configurations used to generate the DGNs. All datasets are provided in a standardized `HDF5` format with fixed training, validation, and test splits, together with the observation masks and ground-truth quantities required for local evaluation. The accompanying repository provides the configurations, dependencies, and evaluation scripts needed to reproduce the benchmark pipeline and reported experiments. Detailed procedures for DGN training and selection, Poisson observation generation, construction of partial-observation conditions, communication-model training and selection, and computation of all benchmark metrics are provided in the Appendix. Together, these resources enable reproduction of the benchmark datasets, model-fitting protocol, and evaluation procedure from the released code and data.

References

Anil Bollimunta and Jochen Ditterich. Local computation of decision-relevant net sensory evidence in parietal cortex. *Cerebral Cortex*, 22(4):903–917, 2012. doi: 10.1093/cercor/bhr165.

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience*, 20(3): 340–352, Mar 2017. ISSN 1546-1726. doi: 10.1038/nn.4497. URL https://doi.org/10.1038/nn.4497.

Carlos D Brody and Timothy D Hanks. Neural underpinnings of the evidence accumulator. *Curr. Opin. Neurobiol.*, 37:149–157, April 2016a.

Carlos D. Brody and Timothy D. Hanks. Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology*, 37:149–157, 2016b. doi: 10.1016/j.conb.2016.01.003.

Susu Chen, Yi Liu, Ziyue Aiden Wang, Jennifer Colonell, Liu D. Liu, Han Hou, Nai-Wen Tien, Tim Wang, Timothy Harris, Shaul Druckmann, Nuo Li, and Karel Svoboda. Brain-wide neural activity underlying memory-guided movement. *Cell*, 187(3):676–691.e16, 2024. doi: 10.1016/j.cell.2023.12.035.

Thomas B. Christophel, P. Christiaan Klink, Bernhard Spitzer, Pieter R. Roelfsema, and John-Dylan Haynes. The distributed nature of working memory. *Trends in Cognitive Sciences*, 21(2):111–124, 2017. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2016.12.007. URL https://www.sciencedirect.com/science/article/pii/S1364661316302170.

Charlie J. Clarke-Williams, Vítor Lopes dos Santos, Laura Lefèvre, Demi Brizee, Adrien A. Causse, Roman Rothaermel, Katja Hartwich, Pavel V. Perestenko, Robert Toth, Colin G. McNamara, Andrew Sharott, and David Dupret. Coordinating brain-distributed network activities in memory resistant to extinction. *Cell*, 187(2):409–427.e19, 2024. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2023.12.018. URL https://www.sciencedirect.com/science/article/pii/S0092867423013508.

Olivier Codol, Jonathan A. Michaels, Mehrdad Kashefi, J. Andrew Pruszynski, and Paul L. Gribble. Motornet, a python toolbox for controlling differentiable biomechanical effectors with artificial neural networks. *eLife*, 12:RP88591, jul 2024. doi: 10.7554/eLife.88591. URL https://doi.org/10.7554/eLife.88591.

Shintaro Funahashi. Working memory in the prefrontal cortex. *Brain Sciences*, 7(5), 2017. ISSN 2076-3425. doi: 10.3390/brainsci7050049. URL https://www.mdpi.com/2076-3425/7/5/49.

Joshua Glaser, Matthew Whiteway, John Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14867–14878. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/aa1f5f73327ba40d47ebce155e785aaf-Paper.pdf.

Evren Gokcen, Anna I Jasper, João D Semedo, Amin Zandvakili, Adam Kohn, Christian K Machens, and Byron M Yu. Disentangling the flow of signals between populations of neurons. *Nat. Comput. Sci.*, 2(8):512–525, August 2022.

Evren Gokcen, Anna Ivic Jasper, Alison Xu, Adam Kohn, Christian K. Machens, and Byron M. Yu. Uncovering motifs of concurrent signaling across multiple neuronal populations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=u39QQh5L8Q.

Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574, 2007. doi: 10.1146/annurev.neuro.29.051605.113038.

Matthew D. Golub and David Sussillo. Fixedpointfinder: A tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks. *Journal of Open Source Software*, 3(31): 1003, 2018. doi: 10.21105/joss.01003. URL https://doi.org/10.21105/joss.01003.

Timothy D. Hanks, Charles D. Kopec, Bingni W. Brunton, Chunyu A. Duan, Jeffrey C. Erlich, and Carlos D. Brody. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*, 520(7546):220–223, 2015. doi: 10.1038/nature14066.

Taishi Iwasaki, Hideitsu Hino, Masami Tatsuno, Shotaro Akaho, and Noboru Murata. Estimation of neural connections from partially observed neural spikes. *Neural Netw.*, 108:172–191, December 2018.

Mitra Javadzadeh and Sonja B. Hofer. Dynamic causal communication channels between neocortical areas. *Neuron*, 110(15):2470–2483.e7, 2022. doi: 10.1016/j.neuron.2022.05.011.

Byungwoo Kang and Shaul Druckmann. Approaches to inferring multi-regional interactions from simultaneous population recordings. *Current Opinion in Neurobiology*, 65:108–119, 2020. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2020.10.004. URL https://www.sciencedirect.com/science/article/pii/S0959438820301422. Whole-brain interactions between neural circuits.

Orren Karniol-Tambour, David M. Zoltowski, E. Mika Diamanti, Lucas Pinto, Carlos D Brody, David W. Tank, and Jonathan W. Pillow. Modeling state-dependent communication between brain regions with switching nonlinear dynamical systems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WQwV7Y8qwa.

Belle Liu, Jacob Sacks, and Matthew D. Golub. Accurate identification of communication between multiple interacting neural populations, 2025. URL https://arxiv.org/abs/2506.19094.

Camden J. MacDowell, Alexandra Libby, Caroline I. Jahn, Sina Tafazoli, and Timothy J. Buschman. Multiplexed subspaces route neural activity across brain-wide networks. *Nature Communications*, 16(1):3359, 2025. doi: 10.1038/s41467-025-58698-2. URL https://doi.org/10.1038/s41467-025-58698-2.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974.

Amin Nejatbakhsh and Yixin Wang. Identifying neural dynamics using interventional state space models, 2025. URL https://openreview.net/forum?id=FwW3jqchtY.

Angela K. Nietz, Laurentiu S. Popa, Martha L. Streng, Russell E. Carter, Suhasa B. Kodandaramaiah, and Timothy J. Ebner. Wide-field calcium imaging of neuronal network dynamics in vivo. *Biology*, 11(11), 2022. ISSN 2079-7737. doi: 10.3390/biology11111601. URL https://www.mdpi.com/2079-7737/11/11/1601.

Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis, 2023. URL https://arxiv.org/abs/2306.10168.

Felix Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark M. Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan W. Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. *CoRR*, abs/2109.04463, 2021. URL https://arxiv.org/abs/2109.04463.

Dean A. Pospisil, Max J. Aragon, Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, Philipp Schlegel, Szi-chieh Yu, Claire E. McKellar, Marta Costa, Katharina Eichler, Gregory S. X. E. Jefferis, Mala Murthy, and Jonathan W. Pillow. The fly connectome reveals a path to the effectome. *Nature*, 634(8032):201–209, 2024. doi: 10.1038/s41586-024-07982-0. URL https://doi.org/10.1038/s41586-024-07982-0.

William Qian, Jacob A. Zavatone-Veth, Benjamin S. Ruben, and Cengiz Pehlevan. Partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics. *bioRxiv*, 2024. doi: 10.1101/2024.05.24.595741. URL https://www.biorxiv.org/content/early/2024/05/26/2024.05.24.595741.

Ehsan Rezayat, Kelsey Clark, Mohammad-Reza A Dehaqani, and Behrad Noudoost. Dependence of working memory on coordinated activity across brain areas. *Front. Syst. Neurosci.*, 15:787316, 2021.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, 2019. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2019.01.026. URL https://www.sciencedirect.com/science/article/pii/S0896627319300534.

Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, December 2019.

David Sussillo, Rafal Jozefowicz, L. F. Abbott, and Chethan Pandarinath. Lfads - latent factor analysis via dynamical systems, 2016. URL https://arxiv.org/abs/1608.06315.

Masaki Takeda, Toshiyuki Hirabayashi, Yusuke Adachi, and Yasushi Miyashita. Dynamic laminar rerouting of inter-areal mnemonic signal by cognitive operations in primate temporal cortex. *Nature Communications*, 9(1):4629, 2018. doi: 10.1038/s41467-018-07007-1.

Eric M. Trautmann, Janis K. Hesse, Gabriel M. Stine, Ruobing Xia, Shude Zhu, Daniel J. O'Shea, Bill Karsh, Jennifer Colonell, Frank F. Lanfranchi, Saurabh Vyas, Andrew Zimnik, Natalie A. Steinmann, Daniel A. Wagenaar, Alexandru Andrei, Carolina Mora Lopez, John O'Callaghan, Jan Putzeys, Bogdan C. Raducanu, Marleen Welkenhuysen, Mark Churchland, Tirin Moore, Michael Shadlen, Krishna Shenoy, Doris Tsao, Barundeb Dutta, and Timothy Harris. Large-scale brain-wide neural recording in nonhuman primates. *bioRxiv*, 2023. doi: 10.1101/2023.02.01.526664. URL https://www.biorxiv.org/content/early/2023/02/04/2023.02.01.526664.

Ejike Ugba and Jan Gertheiss. A modification of mcfadden's R^2 for binary and ordinal response models. *Communications for Statistical Applications and Methods*, 30(1):49–63, 2023. doi: 10.29220/CSAM.2023.30.1.049.

Christopher Versteeg, Jonathan D. McCart, Mitchell Ostrow, David M. Zoltowski, Clayton B. Washington, Laura Driscoll, Olivier Codol, Jonathan A. Michaels, Scott W. Linderman, David Sussillo, and Chethan Pandarinath. Computation-through-dynamics toolkit: Simulated datasets and quality metrics for dynamical models of neural activity. *bioRxiv*, 2026. doi: 10.1101/2025.02.07.637062. URL https://www.biorxiv.org/content/early/2026/06/05/2025.02.07.637062.

Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020. doi: 10.1146/annurev-neuro-092619-094115. URL https://doi.org/10.1146/annurev-neuro-092619-094115.

Xiao-Jing Wang. Theory of the multiregional neocortex: Large-scale neural dynamics and distributed cognition. *Annual Review of Neuroscience*, 45:533–560, 2022. doi: 10.1146/annurev-neuro-110920-035434.

Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.*, 22(2):297–306, February 2019.

A DGN TRAINING DETAILS

All DGN hyperparameters and training configurations are provided in `configs/` in the repository. Across all three DGN classes, we train multiple candidate networks using a random search over discrete hyperparameter grids, varying the number of hidden units per region and the magnitude of dynamic noise. For each candidate DGN, we train a linear decoder on 80% of trials and evaluate decoding performance on the remaining 20% of held-out trials.

We retain the network that achieved the lowest validation loss, but only if the task-relevant ground-truth variables remain linearly decodable from the Poisson spike counts. This procedure enables us to exclude potential networks in which dynamic noise or the Poisson sampling process substantially obscures information represented by the hidden unit activity, which would make communication model recovery results more difficult to interpret.

A.1 DELAYED-MEMORY

The DELAYED-MEMORY DGN contains three regions with 512 hidden units per region and dynamic noise $\boldsymbol{\epsilon}_t^i \sim \mathcal{N}(0, 0.5)$. The objective function includes two mean-squared-error (MSE) losses: one between each region's private input s_t^j and its outgoing messages $\mathbf{m}_t^{j\to i}$ to all target regions A^i, with a two-time-step delay, and another between a readout of each region's hidden state, $\mathbf{y}^i = W_{hy}\mathbf{h}_t^i$, and its received inputs from the preceding two time steps. For network selection, we verify that the relevant private and delayed input signals remain linearly decodable from the corresponding neural populations after the Poisson sampling process.

A.2 RELAY-DECISION

The RELAY-DECISION DGN contains two regions with 256 hidden units per region and dynamic noise $\boldsymbol{\epsilon}_t^i \sim \mathcal{N}(0, 0.001)$. The objective function includes two losses: an MSE loss between a readout of the relay region's hidden state, $\mathbf{y}^R = W_{hy}^R\mathbf{h}_t^R$, and the instantaneous input s_t, and a binary cross-entropy loss between a readout of the decision region's hidden state $\mathbf{y}^D = W_{hy}^D\mathbf{h}_t^D$ and the sign of the cumulative input, $\text{sign}(d_t)$. For network selection, we verify that the instantaneous stimulus s_t remains linearly decodable from the relay-region activity and that the decision variable d_t remains linearly decodable from the decision-region activity after the Poisson sampling process.

A.3 COGNITIVE TASK SUITE.

Each COGNITIVE TASK SUITE DGN contains four regions of 64 hidden units. Unlike the fixed noise levels used for DELAYED-MEMORY and RELAY-DECISION, the scale of dynamic noise ϵ here is gradually increased to 0.1 over the first 200 training epochs to improve training stability while the task-dependent dynamics and communication structure are learned. The objective function includes binary cross-entropy on the saccade output, as well as cross-entropy on the response direction discretized into 36 angular bins.

For each candidate DGN, each communication pathway is sampled independently with connection probability $p = 0.7$, producing relatively dense networks while varying the available connectivity pathways. We reject sampled graphs if any region has zero in-degree or zero out-degree, or if any region cannot reach the output region A^{out} within two hops.

During training, we additionally apply L_2 regularization to both inter-regional communication and local-region recurrent dynamics. For each of the 20 tasks adapted from Yang et al. (2019), we grid search across random seeds, as well as communication and recurrent regularization strengths from $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. Both regularization coefficients are ramped from 30% of their target value to their full value over the first 200 epochs, and early stopping is disabled during this ramp period. We retain only candidate DGNs that successfully perform their assigned task, requiring training accuracy greater than 0.80 and held-out accuracy greater than 0.65.

B Poisson Observation Generation Pipeline

For each dataset, we provide both continuous DGN activity and a sampled Poisson spike-count variant. We sample the Poisson spike counts using the full dataset before constructing train/val/test sets such that all sets share the same normalization statistics and activity-to-rate transformation.

The regional activity $\mathbf{h}$ is first standardized to have zero mean and unit variance:

$$\tilde{\mathbf{h}} = \frac{\mathbf{h} - \mu_{\mathbf{h}}}{\sigma_{\mathbf{h}}},$$

where μ_h and σ_h denote the empirical mean and standard deviation of h, respectively.

The standardized activity is then converted to a firing rate according to

$$\mathbf{r} = r_{\text{scale}} \frac{\tilde{\mathbf{h}} + 1}{2},$$

where we use $r_{\text{scale}} = 40\,\text{Hz}$. Under this mapping, standardized activity values of -1, 0, and 1 correspond to firing rates of 0, 20, and $40\,\text{Hz}$, respectively.

For each time bin of width $\Delta t = 0.01\text{s}$, we compute the Poisson parameter as

$$\lambda = \max(0, \mathbf{r}\Delta t)\,,$$

where the nonnegativity constraint ensures a valid Poisson parameter λ. Spike counts are then sampled independently for each time bin as

$$y_t \sim \text{Poisson}(\lambda).$$

This procedure preserves the underlying DGN dynamics, task structure, and inter-regional communication while introducing stochastic spike-count observations.

C Partial Neuronal and Regional Observability

Partial neuronal observability. For each dataset, we provide four partial neuronal observation masks corresponding to observed-neuron fractions $p \in \{0.25, 0.50, 0.75, 1.00\}$. For a mask with fraction p, a proportion p of the neurons in each region is made available to the communication model, while the remaining $1 - p$ is withheld. The full DGN is simulated before masking, so withheld neurons remain part of the underlying system and can continue to influence the observed activity.

Partial regional observability. For each dataset in the Cognitive Task Suite, we additionally provide two partial regional observation masks, which withhold all neural activity from either A^{fix} or A^{stim}. As with partial neuronal observability, masking is applied only after simulation of the full DGN, so the withheld region remains active and its effects can remain reflected in the observed regions.

D McFadden's Pseudo-R^2 for Poisson Observations

For Poisson spike-count datasets, we report McFadden's pseudo-R^2, a likelihood-based goodness-of-fit measure comparing the submitted model to a null model that predicts the mean spike count (McFadden, 1974; Ugba & Gertheiss, 2023). We define

$$R^2_{\text{McFadden}} = 1 - \frac{\text{NLL}_{\text{model}}}{\text{NLL}_{\text{null}}},$$

where $\text{NLL}_{\text{model}}$ is the Poisson negative log-likelihood of model reconstructed rates and NLL_{null} is the Poisson negative log-likelihood of the mean across trial and time.

E ADDITIONAL BENCHMARK DETAILS

Dataset splits. Each dataset is released with fixed training, validation, and test sets constructed using an 80/10/10 partition, containing 820, 102, and 102 trials, respectively. The training and validation sets are available for model development, while the test set should be reserved for final benchmark evaluation. All sets are provided in a standardized `.h5` format and are derived from the same underlying DGN.

HDF5 data format. Each dataset is stored as a separate `.h5` file. Regional neural activity is stored in datasets of the form `area-<region>`, with shape $K \times T \times n^i$, where K is the number of trials in the split, T is the number of time steps per trial, and n^i is the number of neurons in region i.

The three regions in DELAYED-MEMORY are stored as `area-A0`, `area-A1`, and `area-A2`. The two regions in RELAY-DECISION are stored as `area-R` and `area-D` for the relay and decision regions, respectively. Lastly, the four regions in each COGNITIVE TASK SUITE dataset are stored as `area-A0`, `area-A1`, `area-A2`, and `area-A3`, for A^{fix}, A^{stim}, A^{i}, and A^{output}, respectively.

The released training and validation files only contain regional hidden-state activity, corresponding to the only quantity a communication model should have access to. Ground-truth quantities describing the data-generating system, including inter-regional communication and external inputs are withheld from these sets and reserved for benchmark evaluation. This prevents models from directly accessing the interactions they are intended to infer, ensuring that performance reflects recovery from neural activity rather than direct supervision from the ground truth.

In the test set, we provide the ground-truth communication quantities required to run the benchmark locally. This is only for local evaluation, and models should not use the ground-truth quantities to train, to ensure faithful analyses. Ground-truth inter-regional communication is stored in datasets of the form `message-mesgs` and has shape $K \times T \times d_m$, where d_m is the dimensionality of the communication signal. The RELAY-DECISION dataset's inter-regional communication message is stored in `message-r-to-d`. We additionally include the ground-truth input signals to evaluate external input recovery, which is stored in `truth-inp` for the DELAYED-MEMORY and RELAY-DECISION datasets. For COGNITIVE TASK SUITE datasets, we include ground-truth input signals in `truth-fix` and in `truth-stim1/stim2`.

Partial neuronal observation conditions are specified through indices stored under the `observation/` group of the training file. For each dataset, we provide observation conditions corresponding to 25%, 50%, 75%, and 100% neuronal coverage. Each condition is stored under `observation/pct_<p>/`, where $p \in 25, 50, 75, 100$. Within each observation condition, `observed_idx/<region>` contains the indices of neurons that can be observed by the communication model to be trained, along with `held_out_idx/<region>` that specifies which 10% of neurons outside the observed population is reserved for held-out neuron activity reconstruction evaluation. However, even at 100% observation, we still hold out 10% of neurons to enable accurate held-out evaluation metrics. For COGNITIVE TASK SUITE datasets, we additionally provide partial regional observation masks in which all neural activity from either A^{fix}(`area-A0`) or A^{stim} (`area-A1`) is withheld from the communication model.

Submission format. To run the benchmark locally, users should clone the IRCB '26 repository, install the required dependencies, and provide the corresponding `test.h5` file together with a model-generated prediction file that follows the HDF5 conventions described above. The prediction file should contain reconstructed neural activity for each modeled region using the same trial and time dimensions as the corresponding test data, along with inferred inter-regional communication. Models that explicitly infer external inputs should additionally provide those inferred input signals. The benchmark evaluation scripts compare these submitted quantities against the ground-truth values stored in `test.h5` to compute the applicable reconstruction and communication-recovery metrics.

F ADDITIONAL COGNITIVE TASK SUITE DETAILS

The COGNITIVE TASK SUITE comprises 20 cognitive tasks adapted from Yang et al. (2019). We preserve the core computations and trial structures of these tasks while expressing them in a common format for the multi-region DGNs. Each trial contains a fixation signal, two stimulus modalities, a target response direction, and a saccade signal. Stimulus directions are defined on a circular domain, and decision-making tasks additionally vary stimulus strengths. Trial epoch durations vary randomly across trials and are scaled to fit a fixed 200-step trial duration.

F.1 GO / ANTI TASK FAMILY

The Go / Anti family contains six tasks that vary both the timing of the required response and whether the network should respond toward or away (Anti) from the presented stimulus. In GO, a stimulus appears during fixation and remains present through the response period. After fixation is released, the network reports the stimulus direction. In RT GO, the network instead responds immediately when the stimulus appears. In DLY GO, the stimulus is presented transiently and followed by a delay period, requiring the stimulus direction to be maintained until the response period. The corresponding ANTI, RT ANTI, and DLY ANTI tasks use the same trial structures but require a response in the direction opposite to the stimulus, $\theta + \pi$. Across these tasks, the stimulus is randomly assigned to one of the two stimulus modalities on each trial.

F.2 DECISION-MAKING TASK FAMILY

The decision-making family contains five tasks in which two directional alternatives are presented simultaneously and the network must select one direction based on their stimulus strengths. In DM 1, the alternatives are presented in the first stimulus modality and the target is the direction with greater strength, while DM 2 uses the same computation with stimuli presented in the second modality. In the context-dependent variants, both modalities contain evidence about the same two directional alternatives. CTXT DM 1 and CTXT DM 2 require the network to select the stronger alternative based only on evidence from the first and second modality, respectively. MULTSEN DM requires evidence from both modalities to be integrated, with the response corresponding to the alternative with the greater combined stimulus strength. This task family tests modality-specific decision-making, context-dependent evidence selection, and multisensory evidence integration.

F.3 DELAYED DECISION-MAKING TASK FAMILY

The delayed decision-making family extends the core decision-making tasks and contains five analogous tasks in which the two directional alternatives are presented sequentially rather than simultaneously. DLY DM 1 and DLY DM 2 requires comparison of the sequential alternatives presented in the first and second stimulus modalities, respectively. In CTXT DLY DM 1 and CTXT DLY DM 2, evidence is provided in both modalities, but the decision must be based only on the first or second modality, respectively. DLY MULTSEN DM requires evidence to be integrated across both modalities before selecting the stronger alternative. These tasks jointly require maintenance of stimulus information across time and modality-dependent selection or integration of evidence.

F.4 MATCH / NON-MATCH TASK FAMILY

The match / non-match family contains four tasks in which two stimuli are presented sequentially with an intervening delay and the network determines whether they satisfy a specified matching rule. In DMS (delayed match-to-sample), a go response is required when the two stimuli have the same direction, whereas DNMS (delayed non-match-to-sample) requires a go response when their directions differ. DMC (delayed match-to-category) defines two broad stimulus categories over the circular stimulus space and requires a go response when both stimuli belong to the same category. DNMC reverses this criterion and requires a response when the stimuli belong to different categories. On go trials, the network reports the direction of the second stimulus and produces a saccade signal. On no-go trials, it maintains fixation and produces no directional response.

G Model Training and Validation

G.1 Reduced Rank Regression

Following Liu et al. (2025), we use ridge-regularized reduced-rank regression to linearly predict the activity of each target region A^i. We concatenate trials and time such that the target activity is $Y^i \in \mathbb{R}^{KT \times n^i}$, where K is the number of trials, T is the number of time steps, and n^i is the number of observed neurons in region i. The activities of all source regions A^j, $j \neq i$, are concatenated along the feature dimension to form

$$X^{-i} = \left[X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^N\right] \in \mathbb{R}^{KT \times \sum_{j \neq i} n^j}, \tag{2}$$

where $X^j \in \mathbb{R}^{KT \times n^j}$ and N is the number of regions. We apply the rank constraint separately to the coefficient submatrix corresponding to each source region, preserving a source-specific low-rank mapping rather than imposing a shared low-rank subspace across all sources. To do so, we first obtain the ridge-regression coefficients:

$$W^{\text{ridge}} = \arg\min_W ||Y^i - X^{-i}W||_F^2 + \alpha||W||_F^2.$$

We partition W^{ridge} by source region, such that $W^{\text{ridge},j} \in \mathbb{R}^{n^j \times n^i}$. For each source region j, we compute the singular value decomposition of its predicted contribution to the target,

$$X^j W^{\text{ridge},j} = U^j \Sigma^j (V^j)^\top,$$

and retain the first r^j right singular vectors, $V^j_{r^j}$. The reduced-rank coefficient matrix for source region j is then

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

We additionally shift the activity of each source region by a source-specific time delay δ^j, allowing different source regions to have different transmission delays. Thus, for each target region, the model hyperparameters consist of the ridge regularization strength α, a delay δ^j for each source region j, and a rank r^j for each source region j. We optimize these hyperparameters using an iterative coordinate-wise search, treating each δ^j and r^j as a separate hyperparameter. At each step, we hold all other hyperparameters fixed and select the value of a single hyperparameter by cross-validation. We cycle through α, each source-specific delay δ^j, and each source-specific rank r^j until the selected hyperparameters converge.

G.2 Multi-Region Latent Factor Analysis via Dynamical Systems

We fit MR-LFADS (Liu et al., 2025) to the Poisson observations. Initial experiments using an inferred-input dimension of 6 resulted in poor fits to the observed activity. We then experimented with larger inferred-input dimensions, matching the number of neurons in each region, which improved reconstruction performance and was used for the reported models. We used an inferred-input dimension of 64 for the Cognitive Task Suite, 512 for Delayed-Memory, and 256 for Relay-Decision.

Moreover, learning-rate decay and weight decay both destabilized training, so we use a fixed learning rate of 5×10^{-4} with no weight decay. We sweep the inferred-input and communication KL weights over $(3.0, 0.3)$, $(1.0, 0.1)$, $(0.1, 0.01)$, $(0.01, 0.001)$, and $(0.001, 0.0001)$, selecting the pair with the best held-out neural reconstruction performance. The L_2 penalty on generator weights, with scale 10^4, begins at epoch 100 and ramps over 80 epochs. Controller and initial-condition KL penalties begin at epoch 50 and ramp over 200 epochs, while the communication KL penalty begins at epoch 150 and ramps over 100 epochs. We use a dropout rate of 0.3, a batch size of 16, and gradient clipping at 200. Models are trained for 300 to 1000 epochs, but we checkpoint the model with the lowest exponential moving average of the validation reconstruction loss and apply early stopping after 50 epochs without improvement.

For evaluation, we use the posterior mean of the explicit communication variables as the inferred communication, inferred external inputs are taken from the controller outputs, and reconstructed activity is the exponentiated Poisson log-rate, with held-out neurons filled in from the predictor head.

H PER-TASK RESULTS FOR THE COGNITIVE TASK SUITE

Table 2: **Per-task results for the COGNITIVE TASK SUITE.** Rate and spike reconstruction are reported separately for held-in and held-out neurons and averaged across regions. Communication and input recovery are each reported as (fidelity, specificity). Input recovery is omitted for RRR because it does not infer external inputs.

Dataset	Model	Rate Recon.		Spike Recon.		Effectome		Signal Recovery	
		Held-in	Held-out	Held-in	Held-out	Norm	Dynamic	Communication	Input
Go	RRR	-2.013	-0.701	-0.146	-0.039	0.684	0.454	(0.647, 0.733)	–
	MR-LFADS	0.366	0.319	0.151	0.157	0.803	0.673	(0.978, 0.973)	(0.785, 0.006)
RT Go	RRR	-0.360	0.067	-0.100	-0.034	0.815	0.802	(0.877, 0.721)	–
	MR-LFADS	0.809	0.799	0.146	0.147	0.832	0.853	(0.994, 0.977)	(0.366, 0.007)
Dly Go	RRR	0.165	0.398	-0.037	-0.004	0.779	0.787	(0.736, 0.717)	–
	MR-LFADS	0.641	0.832	0.415	0.109	0.860	0.889	(0.963, 0.983)	(0.748, 0.008)
Anti	RRR	-0.908	-0.244	-0.118	-0.020	0.727	0.559	(0.752, 0.758)	–
	MR-LFADS	0.664	0.645	0.139	0.109	0.740	0.752	(0.984, 0.979)	(0.748, 0.005)
RT Anti	RRR	-1.008	-0.273	-0.098	-0.014	0.772	0.679	(0.871, 0.768)	–
	MR-LFADS	0.810	0.822	0.154	0.169	0.679	0.534	(0.996, 0.980)	(0.363, 0.005)
Dly Anti	RRR	0.330	0.560	-0.022	0.013	0.813	0.786	(0.796, 0.637)	–
	MR-LFADS	0.627	0.791	0.356	0.122	0.823	0.736	(0.974, 0.983)	(0.684, 0.008)
DM 1	RRR	-0.134	0.272	-0.106	-0.054	0.765	0.664	(0.803, 0.706)	–
	MR-LFADS	0.462	0.746	0.407	0.102	0.836	0.757	(0.961, 0.934)	(0.785, 0.608)
DM 2	RRR	0.205	0.482	-0.060	-0.012	0.800	0.778	(0.715, 0.649)	–
	MR-LFADS	0.445	0.800	0.407	0.109	0.849	0.830	(0.931, 0.931)	(0.738, 0.589)
Ctxt DM 1	RRR	0.141	0.301	-0.040	-0.013	0.763	0.748	(0.721, 0.724)	–
	MR-LFADS	0.541	0.829	0.272	0.118	0.816	0.750	(0.976, 0.980)	(0.706, 0.010)
Ctxt DM 2	RRR	-0.134	0.282	-0.126	-0.023	0.828	0.816	(0.720, 0.696)	–
	MR-LFADS	0.505	0.789	0.426	0.105	0.842	0.767	(0.975, 0.979)	(0.601, 0.008)
MultSen DM	RRR	0.211	0.437	-0.061	-0.021	0.845	0.797	(0.779, 0.708)	–
	MR-LFADS	0.503	0.769	0.403	0.108	0.761	0.748	(0.966, 0.942)	(0.758, 0.702)
Dly DM 1	RRR	0.094	0.373	-0.063	0.007	0.839	0.819	(0.769, 0.638)	–
	MR-LFADS	0.544	0.720	0.424	0.119	0.850	0.851	(0.940, 0.918)	(0.709, 0.665)
Dly DM 2	RRR	-0.052	0.268	-0.061	-0.002	0.819	0.814	(0.845, 0.631)	–
	MR-LFADS	0.649	0.523	0.392	0.088	0.864	0.862	(0.957, 0.918)	(0.742, 0.705)
Ctxt Dly DM 1	RRR	-0.193	0.257	-0.082	-0.004	0.804	0.753	(0.696, 0.678)	–
	MR-LFADS	0.471	0.737	0.245	0.137	0.850	0.699	(0.975, 0.978)	(0.734, 0.011)
Ctxt Dly DM 2	RRR	0.070	0.345	-0.044	-0.028	0.807	0.806	(0.769, 0.667)	–
	MR-LFADS	0.467	0.738	0.242	0.130	0.838	0.714	(0.975, 0.974)	(0.708, 0.010)
Dly MultSen DM	RRR	-0.183	0.163	-0.064	-0.007	0.757	0.714	(0.731, 0.657)	–
	MR-LFADS	0.729	0.388	0.401	0.095	0.747	0.690	(0.947, 0.913)	(0.771, 0.689)
DMS	RRR	-1.477	-0.804	-0.078	0.003	0.833	0.721	(0.661, 0.734)	–
	MR-LFADS	0.232	0.260	0.164	0.133	0.857	0.838	(0.979, 0.969)	(0.741, 0.011)
DNMS	RRR	0.262	0.517	-0.078	-0.036	0.710	0.631	(0.600, 0.601)	–
	MR-LFADS	0.724	0.822	0.124	0.060	0.917	0.907	(0.984, 0.976)	(0.820, 0.007)
DMC	RRR	0.258	0.418	-0.038	-0.015	0.836	0.845	(0.766, 0.664)	–
	MR-LFADS	0.744	0.749	0.167	0.126	0.781	0.731	(0.985, 0.975)	(0.734, 0.009)
DNMC	RRR	0.021	0.278	-0.043	0.017	0.755	0.641	(0.825, 0.707)	–
	MR-LFADS	0.761	0.735	0.168	0.108	0.795	0.818	(0.989, 0.979)	(0.830, 0.011)