Information Flows Reveal Computational Mechanisms of RNNs in Contextual Decision-making

Miles Mahon[1] **and Praveen Venkatesh**[2]
University of Washington and Allen Institute, Seattle, WA
[1]milesjmahon@gmail.com, [2]praveen.venkatesh@alleninstitute.org

Abstract

Understanding the information flow of different task-relevant messages within recurrent circuits is crucial to comprehending how the brain works, and in turn, for diagnosing and treating brain disorders. While several information flow methods have focused on functional connectivity and modalities of communication, we do not yet have a principled approach for understanding what information flows can tell us about the effects of causal interventions. In this paper, we consider a measure called M-information flow, proposed by Venkatesh et al. [1], within an artificial recurrent network trained on a contextual decision-making task studied by Mante et al. [2]. We show that M-information flow recapitulates the dynamics of information integration, showing specialization of individual units, and revealing how context information is incorporated to select the appropriate response without affecting the underlying circuit dynamics. We also show how M-information flow predicts the "behavioral outcome" of causal interventions within the network. This leads us to believe that understanding M-information flow within a recurrent network can inform the design of intervention studies, and in future, of stimulation-based treatments for brain disorders.

1 Introduction

Discerning the flows of different types of information within the brain is central to our understanding of both its function and dysfunction. For instance, the "reward circuit" is a network of regions across the brain responsible for reward-based learning, motivation and pleasure [3]. Its atypical function is believed to be the cause of disorders such as addiction and depression [4, 5]. Diagnosing and treating such disorders requires an *understanding of how information flows* within this circuit, both during normal and atypical function [4]. This understanding should allow us to discern how information about different *"messages"* (such as value, choice, reward-prediction error, etc.) flow individually, so as to contribute to a holistic view of brain function. Crucially, such an understanding of information flow should also provide concrete targets for minimal interventions within the circuit—to change its flows in desired ways, so as to treat disorders.

Recent advances in neurotechnologies have allowed us to overcome the first hurdle in mapping information flows in neural circuits by enabling simultaneous recordings of single-unit activity from multiple brain regions [6–8].These advances call for new methods capable of analyzing hypothesized mechanisms of inter-region communication, and identifying candidate targets for optogenetic interventions to validate these hypotheses in biological networks. Currently, methods such as Granger Causality [9, 10] and Transfer Entropy [11] identify the functional connectivity between regions, while newer ideas such as communication subspaces [12] explain "how" the interaction unfolds. However, these tools are not designed to capture *what* the information being communicated is *about* [1, 13]. Furthermore, it is unclear whether these tools can provide targets for interventions to change the behavior of the network in desired ways.

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

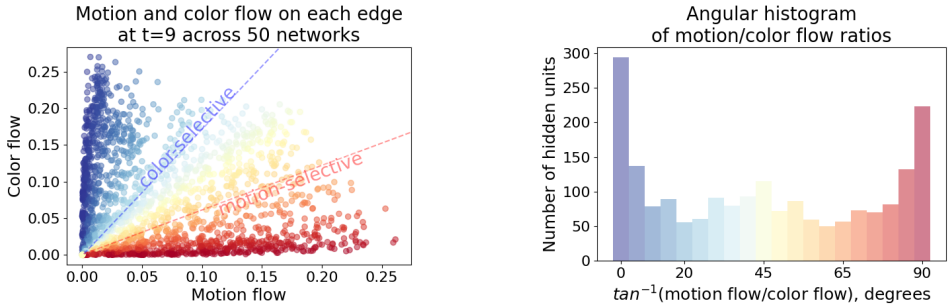

Figure 1: (Left) A scatter plot of color- and motion-information flow for each unit across 50 trained RNNs. (Right) An angular histogram of the inverse tangent of the ratio of motion- to color-information flow for each unit. We find that a majority of units predominantly contain information about only one task.

We explore a recently-developed framework called M-information flow [1], which measures the flow of *specific "messages"* such as stimuli or behavioral variables within a neural circuit. By quantifying information flow using conditional mutual information, this method allows us to *track* the information about any message M as it moves between different neurons or brain regions. Venkatesh et al. [14] showed that this measure is capable of predicting the effect of causal interventions on the edges or nodes of feedforward artificial neural networks in a Fair Machine Learning context. However, the brain is inherently recurrent, so if we wish to understand what M-information flow can tell us about how a circuit computes, or about the effect of a particular intervention, we need to test it on *recurrent* neural networks (RNNs). The nature of recurrent connections also makes understanding interventions more challenging, since (in contrast to feedforward networks) altering one edge affects the information flow at every time step.

2 Methods: The Decision-making Task and Quantifying Information Flow

In this paper, we examine the causal and mechanistic implications of M-information flow by extending it to RNNs trained on a contextual decision-making task.[1] Following Mante et al. [2], we emulate a random-dot paradigm with moving, colored dots, in which the objective is to respond to either the predominant direction of motion or the predominant color based on an independent context signal. The true predominant color, predominant direction of motion and context are all binary signals; color and motion are fed to the network at each time step with additive white Gaussian noise, while context is provided at the penultimate time step (for details, see Appendix A).

We train 50 toy-model RNNs, with four hidden units over 10 time instants, on this contextual decision-making task (details in Appendix A). We measure M-information flow on time-unrolled versions of these RNNs, quantifying the amount of information about the message M that every unit of the RNN has, at each point in time (formally defined in Appendix B). When measuring M-information flow in this task, we take M to be the binary variable representing either the true dominant color or direction of motion; we say "color information flow" to mean the M-information flow of the binary color stimulus.

We examine how information flow recapitulates the RNN's underlying dynamics—in particular, what it tells us about how individual circuit components encode and transmit messages. We also examine whether information flows can be predictive of the effect of certain causal interventions.

3 Results

We measure M-information flow about color and motion in each of the 50 RNNs described above, to make inferences about how the RNNs solve the task: (i) how units specialize for each task, (ii) how flows predict the effect of interventions, and (iii) how the RNNs incorporate information about context in their computations.

Information flow reveals specialization of units for color/motion. We first examine what a visualization of motion- and color-information flow reveals about information processing in our

[1] RNNs have been shown to replicate the dynamics of prefrontal cortex in contextual decision-making [2, 15].

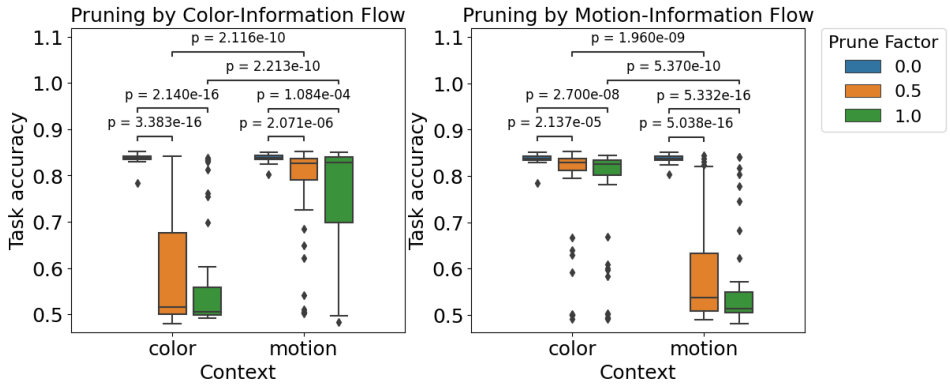

Figure 2: Plots showing the results of pruning the self-edge of nodes with the largest M-information flow about color (left) and motion (right), measured at the last time step $t = 9$. In each plot, the two groups along the x-axis represent the input context, while the y-axis describes the resulting accuracy on held-out data. The box plots describe the variation in the effect of pruning across 50 RNNs, and statistical significance for comparisons is evaluated using a two-sided Mann-Whitney test. The prune factor indicates the extent to which the weight of the self-edge was reduced: 0, 0.5 and 1 indicate a 0%, 50% and 100% reduction in edge weight, respectively. When pruning on the basis of color-information flow (left), we observe a large reduction in accuracy on the color task, with a minimal reduction in accuracy in the motion task. Although pruning has a statistically significant effect on the accuracy on the motion context, the effect size remains small for the median network. Similarly, pruning on the basis of motion-information reduces the accuracy for the motion task to near-chance levels, while the median network faces minimal disruption on the color task.

RNNs. We observe that information about both messages increases over time, indicating integration of both input signals (Fig. 4). We also observe that hidden units appear to specialize, carrying information about either color or motion, but seldom both. Fig. 1(left) shows the color- and motion-information flows at the last time step, for every unit across all 50 RNNs. Fig. 1(right) shows an angular histogram of the same units, using the angle from the x-axis in Fig. 1(left). We find that a majority (74%) of units specialize, and are selective for only one of color- or motion-information. The remainder show mixed selectivity, as seen in shades of yellow in Figs. 1(left) and 1(right), and contain a similar amount of information about both stimuli. We revisit these in the following section.

Information flows predict the effect of interventions. Next, we ask whether M-information flow has causal implications: e.g., can we use M-information flow to design interventions that disrupt the flow of one message but not the other? We examine the effect of interventions on the 50 RNNs trained and measured in the previous section. We hypothesized that highly selective units, if removed, would result in a loss in accuracy for the task about which they convey information, but not the other. This is achieved by intervening on the self-edges of hidden units—i.e., a unit's recurrent connection with itself across time, hindering its ability to integrate information.

We find the units in each network with the highest motion- and color-information flow at the last time step, and reduce the weight of ("prune") their self-edges by half or to zero. When we prune the self-edge of units with the largest color-information flow, we see a large reduction in accuracy on the color task but not the motion task (Fig. 2, left). Conversely, when we prune based on motion-information flow, we see a reduction in motion- but not color-accuracy (Fig 2, right).

Mixed-selective units show a variety of responses to interventions, and the presence or magnitude of motion- or color-information flow do not themselves predict the effect of interventions. Rather, we find that the presence of *redundant* information about M (measured as the *proportion* of M-information flow at a particular unit, out of the total M-information flow over all units at that time step) is correlated with reduced impact of interventions (Pearson-$r = -0.6$, $p = 1.39 \times 10^{-6}$; see Figs. 5 and 6 along with details in Appendix C).

Context presentation suppresses off-context information. The other important component of this task is the ability of the network to selectively respond to color or motion based on the context signal. We next investigate what information flows reveal about how context is processed by our RNNs. We train networks on a version of the task in which context can be presented at any time step,

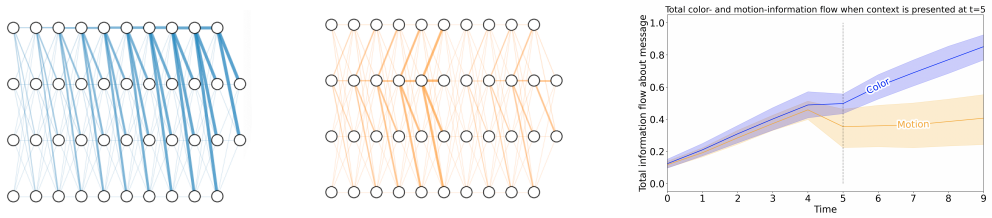

Figure 3: (Left and middle) Information flows about color (left) and motion (middle), in an RNN trained on a version of the task in which context may be presented at any one time step, at random. When the "color" context is shown at the fifth time step, information related to the motion task (shown in orange) is instantaneously suppressed. Color-information flow (shown in blue) continues to integrate as normal. (Right) The sum over all edges of color- and motion-information flow over all ten time steps. Error bars display the variance across ten trained networks. When the "color" context signal is presented at time step $t = 5$, we see a drop in off-context information.

and measure color- and motion-information flows when context is presented at the fifth time step. We find that when the context signal is presented, the flow of information for the off-context task is instantaneously reduced, although integration of new information continues at a slow rate.

Figure 2 shows presentations of the color context signal over 10 RNNs, wherein we see an instantaneous suppression (DC-shift) of information about motion starting at time step 5. This finding supports previous work [2], in which it was shown how the context signal modifies the RNN's population dynamics. It also clarifies an observation in [2] that both on- and off-context signals continue to be integrated, with the distinction that the off-context information is suppressed.

4 Discussion

Our work shows how M-information flow can reveal certain aspects of computational mechanisms (e.g., specialization) in RNNs. It also provides evidence that M-information flow can be predictive of the result of interventions in artificial neural circuits. Further, it helped us clarify the effect of context on the propagation of signals in the network, showing a suppression in off-context information at the time of context presentation, but not a complete cessation of the integration of this information.

Estimating M-information flows is computationally intensive, and for this reason our analysis was limited to small networks. Future work may examine how well these findings generalize to larger networks, including networks with significantly more mixing of signals at the unit level. Similarly, networks with a large number of messages, messages with dependence relationships, and the choice of non-linearity will all need to be examined more carefully. While our networks are small, previous studies [2, 15, 16] have shown the utility of small, recurrent networks in replicating the dynamics of large, biological networks. Furthermore, M-information flow could be applied to a much larger network of units, including in-vivo spike data, by grouping units into brain regions and measuring the flow between regions using the joint activations of all units in each region (or low-dimensional representations thereof).

Ultimately, our work provides evidence that M-information flow is useful for understanding computation and the effects of interventions in recurrent circuits. This could allow neuroscientists to design better interventional studies, for instance, by making observations of M-information flows and then using these to guide optogenetic experiment design, to validate or reject hypotheses. Furthermore, we believe that with better testing and validation, measures such as M-information flow could in future help diagnose and treat brain disorders. For instance, deep brain stimulation has been demonstrated to be effective in treating depression and addiction [17, 18]. Understanding the flow of information specific to a patient's disorder, and being able to infer targets for causal interventions could allow for more personalized stimulation therapies.

Future work could compare M-information flow against other controls, where pruning is performed based on other criteria. For instance, we could construct a naive baseline by pruning the same number of edges selected at random. Other comparisons could include pruning edges selected based on largest absolute activation for a particular message, or edges with the largest absolute correlation with the message (but ignoring the synergistic benefits of maximizing over conditionals).

References

[1] Praveen Venkatesh, Sanghamitra Dutta, and Pulkit Grover. Information flow in computational systems. *IEEE Transactions on Information Theory*, 66(9):5456–5491, 2020. doi: 10.1109/TIT.2020.2987806.

[2] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, November 2013. ISSN 1476-4687. doi: 10.1038/nature12742. URL https://doi.org/10.1038/nature12742.

[3] Robert G. Lewis, Ermanno Florio, Daniela Punzo, and Emiliana Borrelli. The Brain's Reward System in Health and Disease. *Advances in experimental medicine and biology*, 1344:57–69, 2021. ISSN 0065-2598. doi: 10.1007/978-3-030-81147-1_4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8992377/.

[4] Eric J. Nestler. Role of the Brain's Reward Circuitry in Depression: Transcriptional Mechanisms. *International review of neurobiology*, 124:151–170, 2015. ISSN 0074-7742. doi: 10.1016/bs.irn.2015.07.003. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4690450/.

[5] Benjamin A. Ely, Tram N. B. Nguyen, Russell H. Tobe, Audrey M. Walker, and Vilma Gabbay. Multimodal Investigations of Reward Circuitry and Anhedonia in Adolescent Depression. *Frontiers in Psychiatry*, 12, 2021. ISSN 1664-0640. URL https://www.frontiersin.org/articles/10.3389/fpsyt.2021.678709.

[6] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019. doi: 10.1126/science.aav7893. URL https://www.science.org/doi/abs/10.1126/science.aav7893.

[7] Joshua H. Siegle and Xiaoxuan et al. Jia. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-020-03171-x. URL https://doi.org/10.1038/s41586-020-03171-x.

[8] Anne E. Urai, Brent Doiron, Andrew M. Leifer, and Anne K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 25(1):11–19, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00980-9. URL https://doi.org/10.1038/s41593-021-00980-9.

[9] Clive W J Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[10] Andrea Brovelli, Mingzhou Ding, Anders Ledberg, Yonghong Chen, Richard Nakamura, and Steven L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26):9849–9854, 2004. doi: 10.1073/pnas.0308538101. URL https://www.pnas.org/doi/abs/10.1073/pnas.0308538101.

[11] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000. doi: 10.1103/PhysRevLett.85.461. URL https://link.aps.org/doi/10.1103/PhysRevLett.85.461.

[12] 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.

[13] Praveen Venkatesh and Pulkit Grover. Is the direction of greater Granger causal influence the same as the direction of information flow? In *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 672–679, Sept 2015. doi: 10.1109/ALLERTON.2015.7447069.

[14] Praveen Venkatesh, Sanghamitra Dutta, Neil Mehta, and Pulkit Grover. Can Information Flows Suggest Targets for Interventions in Neural Circuits? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3149–3162. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/18de4beb01f6a17b6e1dfb9813ba6045-Paper.pdf.

[15] Michael Kleinman, Chandramouli Chandrasekaran, and Jonathan Kao. A mechanistic multi-area recurrent network model of decision-making. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23152–23165. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c2f599841f21aaefeeabd2a60ef7bfe8-Paper.pdf.

[16] Matthew G. Perich, Charlotte Arlt, Sofia Soares, Megan E. Young, Clayton P. Mosher, Juri Minxha, Eugene Carter, Ueli Rutishauser, Peter H. Rudebeck, Christopher D. Harvey, and Kanaka Rajan. Inferring brain-wide interactions using data-constrained recurrent neural network models. preprint, Neuroscience, December 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.18.423348.

[17] J. Luigjes, W. van den Brink, M. Feenstra, P. van den Munckhof, P. R. Schuurman, R. Schippers, A. Mazaheri, T. J. De Vries, and D. Denys. Deep brain stimulation in addiction: a review of potential brain targets. *Molecular Psychiatry*, 17(6):572–583, June 2012. ISSN 1476-5578. doi: 10.1038/mp.2011.114.

[18] Sibylle Delaloye and Paul E Holtzheimer. Deep brain stimulation in the treatment of depression. *Dialogues in clinical neuroscience*, 2022.

[19] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

[20] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[21] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

A Methods

We train 50 standard RNNs implemented in PyTorch with three inputs, a single layer of four hidden units (with tanh activation functions), and a one-hot encoded, fully-connected, linear output layer of two nodes. Layers (input-hidden, hidden-hidden, hidden-output) are fully-connected. The networks are trained on 10,000 trials (composed of 10 timesteps per trial) over 3,000 epochs. Information flows are estimated on 10,000 held-out datapoints. For all datapoints, context is provided at the last timestep, chosen uniformly at random as ± 1 (and 0, null context, at all other timesteps), corresponding to one of the two independent distributions. In cases where context is presented at a timestep other than the last, such as in Fig. 3, the context signal is presented at a single timestep and is 0 at all others.

The input is composed of three components: a sample drawn from each of two normal distributions, $G^m \overset{iid}{\sim} \mathcal{N}(\mu_m, \sigma_m^2)$ and $G^c \overset{iid}{\sim} \mathcal{N}(\mu_c, \sigma_c^2)$, representing the "motion" and "color" information, respectively. The task of the RNN is to determine either μ_m or μ_c based on the third input (the context signal). The mean of each distribution is chosen independently and with equal probability as ± 1. σ_m^2 and σ_c^2 are defined as $\sqrt{s}/PPF(d)$ where s is the sequence length (10), PPF is the percent point function, and d is chosen uniformly and independently for each distribution over $\{.7, .85, .99\}$, representing the optimal accuracy of the network.

A visualization of a network and its information flows are shown in Figure 4.

A.1 Off-context suppression methods

Fig. 3 was constructed from networks exactly as described in Appendix A, with some exceptions. First, the networks are trained on trials in which context can be presented at any timestep, chosen uniformly at random from all timesteps (i.e., in the range [0,9]). Second, while testing and measuring information flow on held-out test data, the context is presented at a single timestep, $t = 5$, and is 0 at all other timesteps. Finally, the networks use a ReLU activation function, rather than tanh.

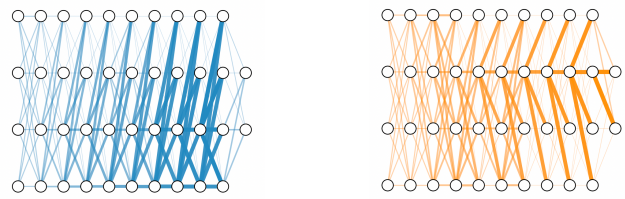

Figure 4: Visualization of an example network, showing the flow of information relating to the mean of the "color" distribution (blue) and of the "motion" distribution (orange). Networks have 4 hidden units and are time-unrolled over 10 timesteps. We find units of the network seem to predominantly hold information about one of the tasks or the other, but not both. We see increasing amounts of information (shown by the thickness of the edges) over time, as new inputs are integrated.

B M-information flow definition

Consider a vanilla RNN, with N_i inputs, N_h hidden units and N_o outputs, with inputs and outputs fully-connected to the hidden units and recurrent connections fully-connected on the hidden units.

Consider a time-unrolled version of this recurrent network, where hidden units are repeated over T time-steps to form a feedforward-like network, as shown in Fig. 4. Let $V_t^{(i)}$ be the i^{th} hidden unit at time t, and let $X(V_t^{(i)})$ represent a random variable describing the distribution of its activation over the input dataset. For a particular message of interest, M (for example, the mean of the motion or color distribution), the M-information flow at unit $V_t^{(i)}$ is given by:

$$F_M(V_t^{(i)}) := \max_{\mathcal{V}'_t \subseteq \mathcal{V}_t \setminus \{V_t^{(i)}\}} I(M; X(V_t^{(i)}) | X(\mathcal{V}'_t)) \tag{1}$$

Where $\mathcal{V}'_t$ represents some subset of all other units at time t (excluding $V_t^{(i)}$), and $I(A; B|C)$ is the conditional mutual information between A and B given C [19]. We take the maximum over all possible subsets [1].

The definition guarantees an unbroken path of information – information may not disappear at some timestep and then reappear at some later timestep [1]. Similarly, in a time-unrolled recurrent network, the method is able to provide a dynamic understanding of how the information in the network changes over time, based on new inputs and new communication.

B.1 Estimating M-information flow

We estimate M-information flow using the same methods proposed in [14, App. A.3]. The method relies on an approximation of mutual information from correlation, which is known to be exact for Gaussian distributions. For non-Gaussian variables, this method may have inaccuracies, but has the benefit of being extremely fast, and thus scaling to larger networks. Future work could consider the application of more accurate mutual information estimators (e.g., [20, 21]).

C Information flow-ratio predicts effect of interventions on mixed-selective units.

While most of the hidden units in the measured networks show strong specialization for one or the other task (just 53 of the 200 units have a flow ratio between 0.6 and 1.7, or 30 and 60 degrees in Fig. 1), and pruning the unit with the largest flow of a particular message predictably reduces the corresponding task accuracy on average (Fig. 2), we wanted to investigate the effect of pruning mixed-selective units. While we find a strong correlation between the information flow on a unit and the disruption of task accuracy upon pruning the unit, we did not observe as clear of a relationship in units with similar amounts of information about both messages (Figs. 5, 6). The reduction in task accuracy from pruning a particular mixed-selective unit and that unit's ratio of information about one task to information about the other are less clearly related at the level of particular networks (Fig. 5).

However, looking over a slightly larger set of mixed-selective units, we observe a relationship between the reduction in task accuracy and the within-task information ratio of the unit. That is, units containing a higher proportion of the total information in the network about a task at the same timestep ($F_M(V_t^{(i)}) / \sum_{j \neq i} F_M(V_t^{(i)})$) have a larger effect on the task's accuracy (fig. 5). This suggests redundancy, and information decomposition more generally, could play a role in future studies of targeted interventions in networks when selectivity is more mixed. In fact, pruning networks based on the information ratio (Fig. 6) has similarly predictable results to pruning the maximum single flow magnitude unit (Fig. 2), though the effect is less pronounced.

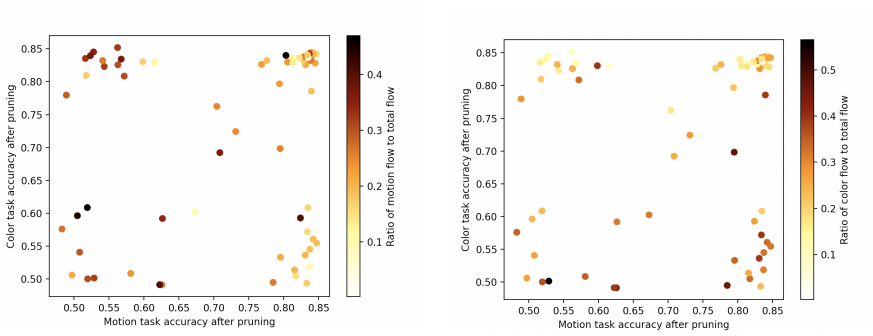

Figure 5: (a) Units with a ratio of motion to color flow between 0.6 and 1.7, colored by the percentage of total motion information in the network flowing to the output. We observe the units with the highest percentage of total information contained in their activations cause the largest drop in motion task accuracy when pruned. (b) The units colored by ratio of color information in the unit to the total information. We see the same effect, clusters of units causing large drops in color task accuracy contain a larger percentage of the total color information in the network.

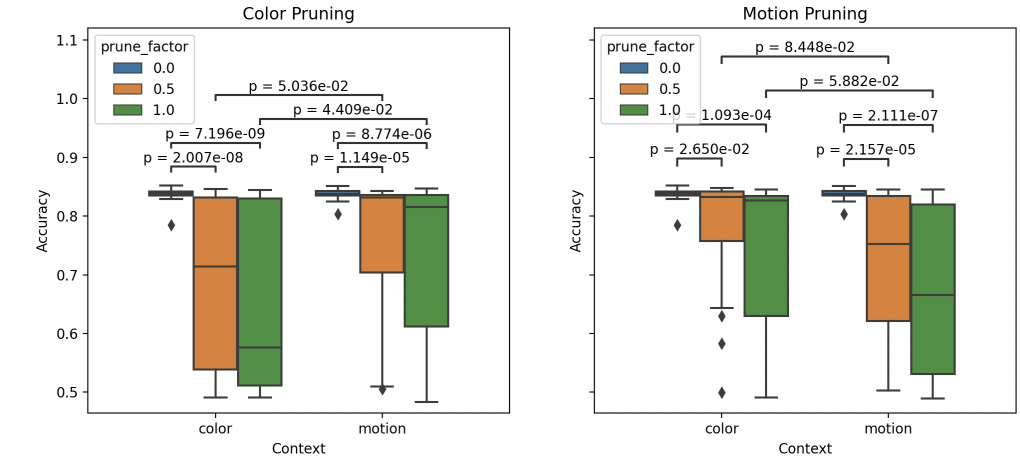

Figure 6: Units in which a mixed-selective unit is pruned. On the left, units removed have a larger color flow ratio: that is, the color flow of the unit proportional to the total color flow to the output, is higher than the motion flow ratio. On the right, units removed have a higher motion flow ratio.