

# Closed-Loop Identifiability in Neural Circuits

true

## Abstract

!!!! Todo 3/16: - Mention basic science applications of CL control - Maybe more forecasting idea of shaping correlations? (don't want reader to be surprised by structure of paper's argument)

==testing==

$$x = 2^2$$

The necessity of intervention in inferring cause has long been understood in neuroscience. Recent work has highlighted the limitations of passive observation and single-site lesion studies in accurately recovering causal circuit structure. The advent of optogenetics has facilitated increasingly precise forms of intervention including closed-loop control which may help eliminate confounding influences. However, it is not yet clear how best to apply closed-loop control to leverage this increased inferential power. In this paper, we use tools from causal inference, control theory, and neuroscience to show when and how closed-loop interventions can more effectively reveal causal relationships. We also examine the performance of standard network inference procedures in simulated spiking networks under passive, open-loop and closed-loop conditions. We demonstrate a unique capacity of feedback control to distinguish competing circuit hypotheses by disrupting connections which would otherwise result in equivalent patterns of correlation<sup>1</sup>. Our results build toward a practical framework to improve design of neuroscience experiments to answer causal questions about neural circuits.

## Introduction

### Estimating causal interactions in the brain

!!!! - 70% done

!!!! Todo 3/16: - "We first propose..." paragraph (could build out or move or change focus away from the 'framework') - think about condensing and/or moving "Inferring causal interactions from time series" subsection - Maybe add half a paragraph or so in the discussion about how causal inference tools can help above correlation analysis (e.g., PC algorithm)

Many hypotheses about neural circuits are phrased in terms of causal relationships: "will changes in activity to this region of the brain produce corresponding changes in another

<sup>1</sup>may end up discussing quantitative advantages such as bidirectional variance (and correlation) control. If that's a strong focus in the results, should be talked about more in the abstract also

region?" Understanding these causal relationships is critical to both scientific understanding and to developing effective therapeutic interventions, which require knowledge of how potential therapies will impact brain activity and patient outcomes.

A range of mathematical and practical challenges make it difficult to determine these causal relationships. In studies that rely only on observational data, it is often impossible to determine whether observed patterns of activity are caused by known and controlled inputs, or whether they are instead spurious connections generated by recurrent activity, indirect relationships, or unobserved "confounders." It is generally understood that moving from experiments involving passive observation to more complex levels of intervention allows experimenters to better tackle challenges to circuit identification. However, while chemical and surgical lesion experiments have historically been employed to remove the influence of possible confounds, they are likely to dramatically disrupt circuits from their typical functions, making conclusions about underlying causal structure drawn from these experiments unlikely to hold in naturalistic settings (Chicharro and Ledberg 2012). *Closed-loop* interventions [...] == "Adam-to-Do" (2022): short description of closed-loop in neuro, maybe drawing from text in this collapsable: ==

Proposal text to draw from:

For decades, engineers have used feedback control to actuate a system based on measured activity to reduce variability, compensate for imperfect measurements, drive systems to desired set points, and decouple connected systems [...]

There is an increasing interest in using approaches from closed-loop control for neural stimulation to both study complex neural circuits and treat neurologic disorders. Recently, a growing community is developing and applying closed-loop stimulation strategies at the cellular and circuit level (Miranda-Dominguez, Gonía, and Netoff 2010; Santaniello, Burns, et al. 2011; Ching et al. 2013; Iolov, Ditlevsen, and Longtin 2014; Nandi, Kafashan, and Ching 2016; Bolus et al. 2018) to understand the brain (Packer et al. 2015) as well as treat disorders (Santaniello, Fiengo, et al. 2011; Paz et al. 2013; Ehrens, Sritharan, and Sarma 2015; Choi et al. 2016; Yang and Shanechi 2016; Kozák and Berényi 2017; Sorokin et al. 2017) The advent of optogenetic stimulation has accelerated the potential for effective closed-loop stimulation by providing actuation strategies that can be more precisely targeted and have minimal recording artifacts compared to conventional microelectrode stimulation (Grosenick, Marshel, and Deisseroth 2015)

Most applications of closed-loop control to neuroscience to

date have used “activity-guided / responsive / triggered stimulation” wherein a predefined stimulus is delivered in response to a detected event. For example, in (Krook-Magnuson et al. 2013) the authors detect seizure activity from spiking and local field potential features to trigger a pulse-train of inhibitory optogenetic stimulation which interrupts the seizure. While this is an effective approach for many applications, these types of closed-loop experiments should be distinguished from closed-loop with ongoing feedback such as dynamic clamp. In these feedback control approaches parameters of stimulation are adjusted on much faster timescales in response to measured activity. For dynamic clamp experiments, this low-latency ongoing feedback control allows experimenters to deliver currents which mimic virtual ion channels which would be implausible with triggered predefined stimulation. These approaches provide additional precision in being able to drive activity patterns, but also come with increased algorithmic and hardware demands. For the rest of this document, we will use “closed-loop control” or “feedback control” to refer to this second, more specific class of approaches.

While many such new actuation and measurement tools have recently become available for neural systems, we require the development of principled algorithmic tools for designing feedback controllers to use these neural interfaces. Our collaborators have previously demonstrated successful closed-loop optogenetic control (CLOC) in-vitro (Newman et al. 2015) and in-vivo (Bolus et al. 2018) to track naturalistic, time-varying trajectories of firing rate.

□ Also add citation to [?]

Despite the promise of these closed-loop strategies for identifying causal relations in neural circuits, however, it is not yet fully understood *when* more complex intervention strategies can provide additional inferential power, or *how* these experiments should be optimally designed. In this paper we demonstrate when and how closed-loop interventions can reveal the causal structure governing neural circuits. Drawing from ideas in causal inference (Pearl 2009) (Maathuis and Nandy 2016) [?], we describe the classes of models that can be distinguished by a given set of input-output experiments, and what experiments are necessary to uniquely determine specific causal relationships.

We first propose a mathematical framework that describes how open- and closed-loop interventions impact observable qualities of neural circuits. Using this framework, experimentalists propose a set of candidate hypotheses describing the potential causal structure of the circuit under study, and then select a series of interventions that best allows them to distinguish between these hypotheses. Using both simple controlled models and in silico models of spiking networks, we explore factors that govern the efficacy of these types of interventions. Guided by the results of this exploration, we present a set of recommendations that can guide the design of open- and closed-loop experiments to better uncover the causal structure underlying neural circuits.

**Inferring causal interactions from time series.** A

number of strategies have been proposed to detect causal relationships between observed variables. Wiener-Granger (or predictive) causality states that a variable  $X$  “Granger-causes”  $Y$  if  $X$  contains information relevant to  $Y$  that is not contained in  $Y$  itself or any other variable [?]. This concept has traditionally been operationalized with vector autoregressive models [?]; the requirement that *all* potentially causative variables be considered makes these notions of dependence susceptible to unobserved confounders [?].

Our work initially focuses on measures of directional interaction that are based on lagged correlations [?]. These metrics look at the correlation of time series collected from pairs of nodes at various lags and detect peaks at negative time lags. Such peaks could indicate the presence of a direct causal relationship – but they could also stem from indirect causal links or hidden confounders [?]. In these bivariate correlation methods, it is thus necessary to consider patterns of correlation between many pairs of nodes in order to differentiate between direct, indirect, and confounding relationships [?]. This distinguishes these strategies from some multivariate methods that “control” for the effects of potential confounders. While cross-correlation-based measures are generally limited to detecting linear functional relationships between nodes, their computational feasibility makes them a frequent metric of choice in experimental neuroscience work [?] [?] [?].

Other techniques detect directional interaction stemming from more general or complex relationships. Information-theoretic methods, which use information-based measures to assess the reduction in entropy knowledge of one variable provides about another, are closely related to Granger causality [?] [?]. The *transfer entropy*  $T_{X \rightarrow Y}(t) = I(Y_t: X_{<t} | Y_{<t})$  extends this notion to time series by measuring the amount of information present in  $Y_t$  that is not contained in the past of either  $X$  or  $Y$  (denoted  $X_{<t}$  and  $Y_{<t}$ ) [?]. Using transfer entropy as a measure of causal interaction requires accounting for potential confounding variables; the *conditional transfer entropy*  $T_{X \rightarrow Y|Z}(t) = I(Y_t: X_{<t} | Y_{<t}, Z_{<t})$  conditions on the past of other variables to account for their potential confounding influence [?, Sec. 4.2.3]. Conditional transfer entropy can thus be interpreted as the amount of information present in  $Y$  that is not contained in the past of  $X$ , the past of  $Y$ , or the past of other variables  $Z$ .

To quantify the strength of causal interactions, information-theoretic and transfer-entropy-based methods typically require knowledge of the ground truth causal relationships that exist [?] or an ability to perturb the system [?] [?]. In practice, these quantities are typically interpreted as “information transfer,” and a variety of estimation strategies and methods to automatically select the conditioning set (i.e., the variables and time lags that should be conditioned on) are used (e.g., [?]). Multivariate conditional transfer entropy approaches using various variable selection schemes can differentiate between direct interactions, indirect interactions, and common causes, but their results depend on choices such as the binning

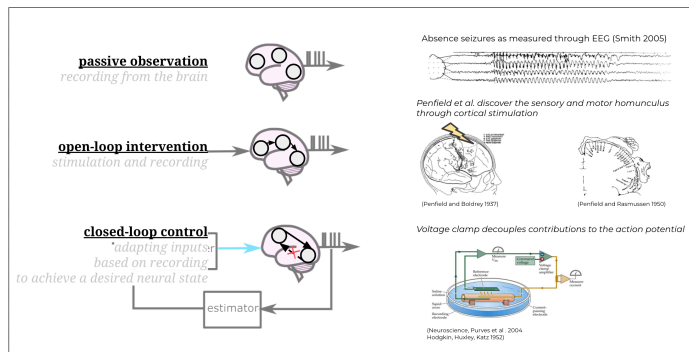
strategies used to discretize continuous signals, the specific statistical tests used, and the estimator used to compute transfer entropy [?]. [If we end up making the jump to IDTx1 in our results: In our empirical results using transfer-entropy-based notions of directional influence we use the IDTx1 toolbox \cite{wollstadt2019idtx1}.] However, despite their mathematical differences, previous work has found that cross-correlation-based metrics and information-based metrics tend to produce qualitatively similar results, with similar patterns of true and false positives [?].

## Interventions in neuroscience & causal inference

!!!! - 70% done

!!!! Todo - Get language more precise and effective (*see writing\_tasks*)

Data collected from experimental settings can provide more inferential power than observational data alone. For example, consider an experimentalist who is considering multiple causal hypotheses for two nodes under study,  $x$  and  $y$ : the hypothesis that  $x$  is driving  $y$ , the hypothesis that  $y$  is driving  $x$ , or the hypothesis that the two variables are being independently driven by a hidden confounder. Observational data revealing that  $x$  and  $y$  produce correlated time-series data is equally consistent with each of these three causal hypotheses, providing the experimentalist with no inferential power. Experimentally manipulating  $x$  and observing the output of  $y$ , however, allows the scientist to begin to establish which causal interaction pattern is at work. Consistent with intuition from neuroscience literature, a rich theoretical literature has described the central role of interventions in inferring causal structure from data [?, ?].



> **Figure INTRO:** Examples of the roles interventions have played in neuroscience. (A) *Passive observation* does not involve stimulating the brain. In this example, passive observational data is used to identify patients suffering from absence seizures. (B) *Open-loop stimulation* involves recording activity in the brain after perturbing a region with a known input signal. Using systematic *open-loop stimulation experiments*, Penfield uncovered the spatial organization of how senses and movement are mapped in the cortex [?] [?]. (C) *Closed-loop control* uses feedback control to precisely specify activity in certain brain regions regardless of activity in other regions. Using closed-loop

control, ==todo-Adam== [?].

The inferential power of interventions is depends on *where* stimulation is applied: interventions on some portions of a system may provide more information about the system's causal structure than interventions in other areas. And interventions are also more valuable when they more effectively set the state of the system: “perfect” closed-loop control, which completely severs a node's activity from its inputs, are often more informative than “soft” interventions that only partially control a part of the system [?].

In experimental neuroscience settings, experimenters are faced with deciding between interventions that differ in both location and effectiveness. For example, stimulation can often only be applied to certain regions of the brain. And while experimenters may be able to exactly manipulate activity in some parts of the brain using closed-loop control, in other locations it may only be possible to apply weaker forms of intervention that perturb a region but do not manipulate its activity exactly to a desired state. In Section [X], we compare the effectiveness of open-loop, closed-loop, and partially-effective closed-loop control.

Although algorithms designed to choose optimal interventions are often designed for simple models with strong assumptions,<sup>2</sup> they provide intuition that can aid practitioners seeking to design real-world experiments that provide as much scientific insight as possible.<sup>3</sup> Importantly, the informativeness of interventions is often independent of the algorithm used to infer causal connections, meaning that certain interventions can reveal portions of a circuit's causal structure that would be impossible for *any* algorithm to infer from only observational data [?] ==(<- Matt to Adam: make sure this citation is in the right place)==. We similarly expect the results we demonstrate in this paper to both inform experimentalists and open avenues for further research.

## Representations & reachability

!!!! - 60% done

!!!! todo - Rewrite  $X=XW+E$  as vector version - Describe what ‘reachability’ is (*see writing\_tasks*)

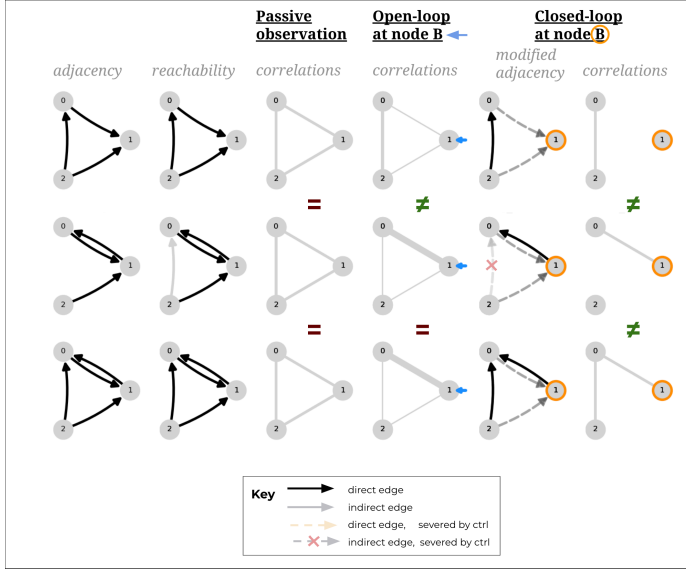
@ import “/section\_content/background\_representation\_reach.md”

!!!! - 70% done

!!!! todo - Talk about what ‘reachability’ means (total direct+indirect impact) - [Matt:] Rewrite first paragraph to not use notation (place this box before any theory/notation sections) - [Matt:] Set expectation here that we’re talking about linear Gaussian circuits

<sup>2</sup>These assumptions are typically on properties such as the types of functional relationships that exist in circuits, the visibility and structure of confounding relationships, and noise statistics.

<sup>3</sup>if citations needed here, could start by looking for a good high-level reference in either [?] or [?]. (Both of these papers are pretty technical, so likely wouldn’t be great citations on their own.)



> **Figure DEMO (box format): Applying CLINC to distinguish a pair of circuits** > Consider the three-node identification problem shown in the figure above, in which the experimenter has identified three hypotheses for the causal structure of the circuit. These circuit hypotheses, shown as directed graphs in column 1, can each also be represented by an adjacency matrix of the form  $W$ : for example, circuit A is represented by an adjacency matrix in which  $w_{01}$ ,  $w_{20}$ , and  $w_{21} \neq 0$ . Note that hypotheses A and C have direct connections between nodes 0 and 2; while hypothesis B does not have a direct connection between these nodes, computing the weighted reachability matrix  $\tilde{W}$  in circuit B an *indirect* connection exists through the path  $2 \rightarrow 1 \rightarrow 0$  (illustrated in gray in column 2). > > Because there are direct or indirect connections between each pair of nodes, passive observation of each hypothesized circuit would reveal that each pair of nodes is correlated (column 3). These three hypotheses are therefore difficult to distinguish<sup>4</sup> for an experimentalist who performs only passive observation, but can be distinguished through stimulation. > > Column 4 shows the impact on observed correlations of performing *open-loop* control on node 1. In hypothesis A, node 1 is not a driver of other nodes, so open-loop stimulation at this site will not increase the correlation between the signal observed at node 1 and other nodes. The path from node 1 to 0 in hypotheses B and C, meanwhile, causes the open-loop stimulation at node 1 to *increase* the observed correlation between nodes 1 and 0. An experimenter can thus distinguish between hypothesis A and the other two hypotheses by applying open-loop control and observing the resulting pattern of correlations (column 4). However, this pattern of open-loop stimulation would not allow the experimenter to distinguish between hypotheses B and C. > > *Closed-loop* control (columns 5 and 6) can provide the experimenter with even more inferential power. Column 5 shows the resulting adjacency matrix when this closed-loop control is applied to node 1. In each hypothesis, the impact of this closed-loop control

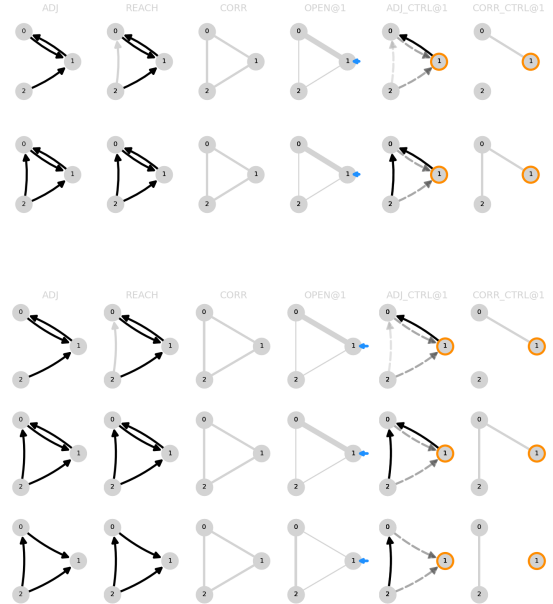
<sup>4</sup>saying “difficult to distinguish” instead of “indistinguishable” here since the magnitudes of the correlations could also be informative with different assumptions

is to remove the impact of other nodes on node 1, because when perfect closed-loop is applied the activity of node 1 is completely independent of other nodes. (These severed connections are depicted in column 5 by dashed lines.) In hypothesis B, this also results in the elimination of the indirect connection from node 2 to node 1. The application of closed-loop control at node 1 thus results in a different observed correlation structure in each of the three circuit hypotheses (column 6). This means that the experimenter can therefore distinguish between these circuit hypotheses by applying closed-loop control – a task not possible with passive observation or open-loop control.

figure to do items for “Adam-to-Do” (2022)

- ☐ “Adam-to-Do” (2022) - change labels at top from “B” to “1”
- ☐ “Adam-to-Do” (2022) - add (A) (B) (C) labels to each row
- ☐ “Adam-to-Do” (2022) - in legend, change in/direct “edge” to in/direct “connection”
- ☐ “Adam-to-Do” (2022) - in legend, orange dashed arrow to dark gray

2,3 circuit versions, straight from code



> 3 circuit walkthrough, walkthrough will all intervention locations might be appropriate for the supplement

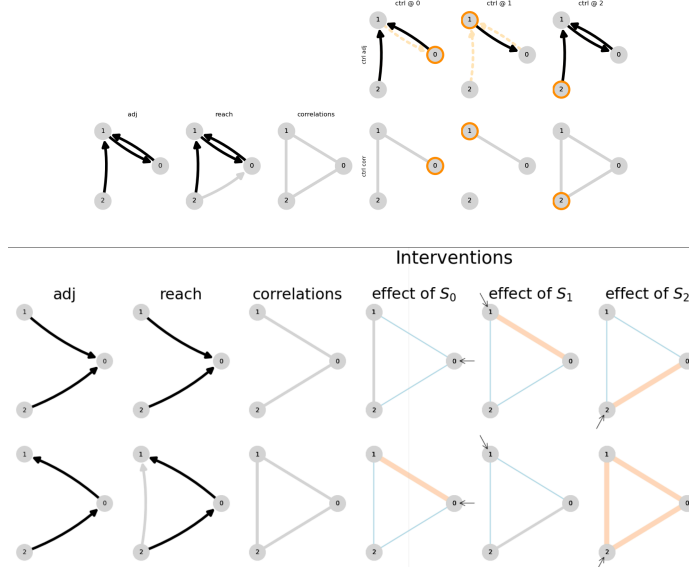
to do items

- ☐ find and include frequent circuit (curto + motif)
- ☐ wrap circuits we want in `example_circuits.py`
- ☐ alt method of displaying indirect paths?
  - <https://networkx.org/documentation/stable/reference/algorithms/>

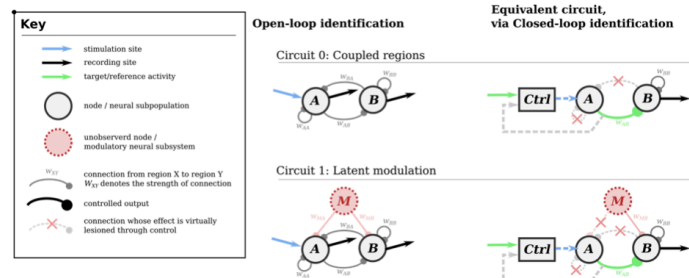
see also

more inspiration: - Combining multiple functional connectivity methods to improve causal inferences - Advancing functional connectivity research from association to causal

tion - Fig1. of “Systematic errors in connectivity”



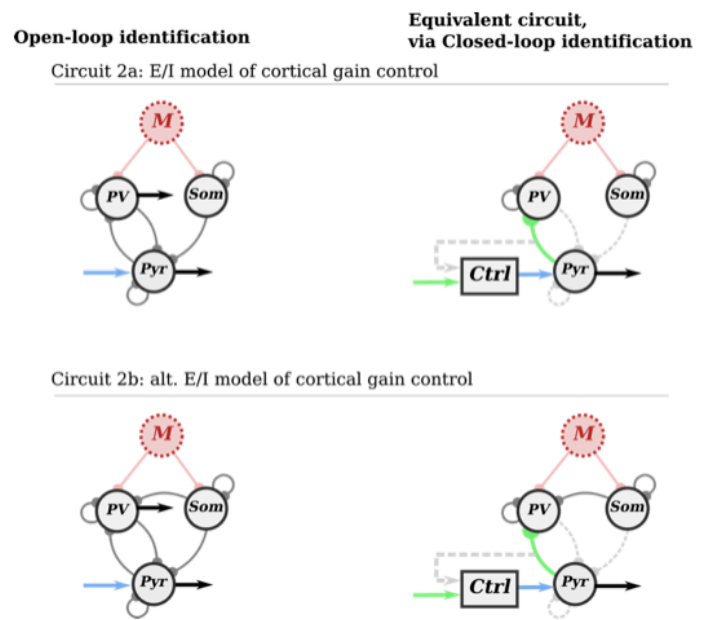
this figure does a great job of: - setting up a key - incrementally adding confounds - highlighting severed edges this figure does NOT - explicitly address multiple hypotheses



**Figure 11: Closed-loop control compensates for inputs to a node in simple circuits:** The left column shows a simple circuit and recording and stimulation sites for an open-loop experiment. The right column shows the functional circuit which results from closed-loop control of the output of region A. Generally, assuming perfectly effective control, the impact of other inputs to a controlled node is nullified and therefore crossed off the functional circuit diagram.

**Figure 11: Closed-loop control compensates for inputs to a node in simple circuits:** The left column shows a simple circuit and recording and stimulation sites for an open-loop experiment. The right column shows the functional circuit which results from closed-loop control of the output of region A. Generally, assuming perfectly effective control, the impact of other inputs to a controlled node is nullified and therefore crossed off the functional circuit diagram.

this figure does a great job of: - using a minimal version of the key above - showing two competing hypotheses - (throughs latent / common modulation in for fun)



**Figure 12: Closed-loop control allows for two circuit hypotheses to be distinguished.** Two hypothesized circuits for the relationships between pyramidal (Pyr, excitatory), parvalbumin-positive (PV, inhibitory), and somatostatin-expressing (Som, inhibitory) cells are shown in the two rows. Dashed lines in the right column represent connections whose effects are compensated for through closed-loop control of the Pyr node. By measuring correlations between recorded regions during closed-loop control it is possible to distinguish which hypothesized circuit better matches the data. Notably in the open-loop intervention, activity in all regions is correlated for both hypothesized circuits leading to ambiguity.

more notes

probably want - two circuits which look clearly different - ! but which have equivalent reachability - possibly with reciprocal connections - possibly with common modulation

- do we need to reflect back from set of possible observations to consistent hypotheses?
  - mention markov equivalence classes explicitly?
- intuitive explanation using binary reachability rules
- point to the rest of the paper as deepening and generalizing these ideas
- (example papers - Advancing functional connectivity research from association to causation, Combining multiple functional connectivity methods to improve causal inferences)
- connect **graded reachability** to ID-SNR
  - $IDSNR_{ij}$  measures the strength of signal related to the connection  $i \rightarrow j$  relative to in the output of node  $j$
  - for true, direct connections this quantity increasing means a (true positive) connection will be identified more easily (with high certainty, requiring

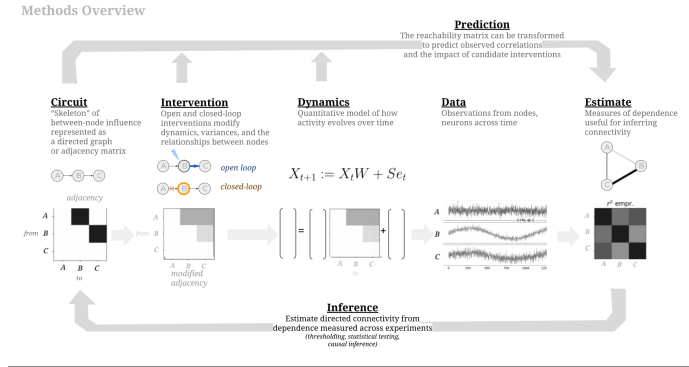


less data)

- for false or indirect connections, this quantity increasing means a false positive connection is more likely to be identified
- as a result we want to maximize IDSNR for true links, and minimize it for false/indirect links

( see also `sketches_and_notation/walkthrough_EI_dissection.md` )

## Theory / Prediction



> **Figure OVERVIEW:** ...

### Predicting correlation structure (theory)

A linear-Gaussian circuit can be described by 1) the variance of the gaussian private (independent) noise at each node, and 2) the weight of the linear relationships between each pair of connected nodes. Let  $s \in \mathbb{R}^p$  denote the variance of each of the  $p$  nodes in the circuit, and  $W \in \mathbb{R}^{p \times p}$  denote the matrix of connection strengths such that

$$W_{ij} = \text{strength of } i \rightarrow j \text{ connection.}$$

Note that  $[(W^T)s]_j$  gives the variance at node  $j$  due to length-1 (direct) connections, and more generally,  $[(W^T)^k s]_j$  gives the variance at node  $j$  due to length- $k$  (indirect) connections. The *total* variance at node  $j$  is thus  $[\sum_{k=0}^{\infty} (W^T)^k s]_j$ .

Our goal is to connect private variances and connection strengths to observed pairwise correlations in the circuit. Defining  $X \in \mathbb{R}^{p \times n}$  as the matrix of  $n$  observations of each node, we have<sup>5</sup>

$$\begin{aligned} \Sigma &= \text{cov}(X) = \mathbb{E}[XX^T] \\ &= (I - W^T)^{-1} \text{diag}(s)(I - W^T)^{-T} \\ &= \widetilde{W} \text{diag}(s) \widetilde{W}^T, \end{aligned}$$

where  $\widetilde{W} = \sum_{k=0}^{\infty} (W^T)^k$  denotes the *weighted reachability matrix*, whose  $(i, j)^{\text{th}}$  entry indicates the total influence of

<sup>5</sup>To see this, denote by  $E \in \mathbb{R}^{p \times n}$  the matrix of  $n$  private noise observations for each node. Note that  $X = W^T X + E$ , so  $X = E(I - W^T)^{-1}$ . The covariance matrix  $\Sigma = \text{cov}(X) = \mathbb{E}[XX^T]$  can then be written as  $\Sigma = \mathbb{E}[(I - W^T)^{-1} E E^T (I - W^T)^{-1}] = (I - W^T)^{-1} \text{cov}(E) (I - W^T)^{-T} = (I - W^T)^{-1} \text{diag}(s) (I - W^T)^{-T}$ .

node  $i$  on node  $j$  through both direct and indirect connections.<sup>6</sup> That is,  $\widetilde{W}_{ij}$  tells us how much variance at node  $j$  would result from injecting a unit of private variance at node  $i$ . We can equivalently write  $\Sigma_{ij} = \sum_{k=1}^p \widetilde{W}_{ik} \widetilde{W}_{jk} s_k$ .

Under passive observation, the squared correlation coefficient can thus be written as

$$\begin{aligned} r^2(i, j) &= \frac{\Sigma_{ij}}{\Sigma_{ii} \Sigma_{jj}} \\ &= \frac{(\sum_{k=1}^p \widetilde{W}_{ik} \widetilde{W}_{jk} s_k)^2}{(\sum_{k=1}^p \widetilde{W}_{ik}^2 s_k) (\sum_{k=1}^p \widetilde{W}_{jk}^2 s_k)}. \end{aligned}$$

This framework also allows us to predict the impact of open- and closed-loop control on the pairwise correlations we expect to observe. To model the application of open-loop control on node  $c$ , we add an arbitrary amount of private variance to  $s_c$ :  $s_c \leftarrow s_c + s_c^{(OL)}$ . To model the application of closed-loop control on node  $c$ , we first sever inputs to node  $c$  by setting  $W_{k,c} = 0$  for  $k = 1, \dots, p$ , and then set the private variance of node  $c$  by setting  $s_c$  to any arbitrary value. Because  $c$ 's inputs have been severed, this private noise will become exactly node  $c$ 's output variance.

!!!! todo [Matt:] add table from `sketches_and_notation/intro-background/causal_vs_expt.md` and modify text above to match

!!!! todo - Some redundancy with simulation methods; cut and paste anything useful in 4.2 and put into 3.1 / 3.2

## Simulation Methods

!!!! todo - reorganize / split sections

@ import “methods0\_0\_overview.md”

### Modeling network structure and dynamics

!!!! - 70% done

to do

- [~] read e.g.
- discuss networks - adj
- discuss 2 key dimensions of complexity
  - linear-gaussian v.s. spiking (LIF - Poisson?)
  - contemporaneous v.s. delayed connections
- discuss brain implementation (supplement)

We sought to understand both general principles (abstracted across particulars of network implementation) as well as some practical considerations introduced by dealing with spikes and synapses.

### Stochastic network dynamics

The first approach is accomplished with a network of nodes with gaussian noise sources, linear interactions, and linear

<sup>6</sup>We can use  $p - 1$  as an upper limit on the sum  $\widetilde{W} = \sum_{k=0}^{\infty} W^k$  when there are no recurrent connections.

dynamics. The second approach is achieved with a network of nodes consisting of populations of leaky integrate-and-fire (LIF) neurons. These differ from the simpler case in their nonlinear-outputs, arising from inclusion of a spiking threshold. Interactions between neurons happen through spiking synapses, meaning information is passed between neurons sparsely in time<sup>7</sup>.

*Neuron dynamics:*

$$\frac{dV}{dt} = \frac{V_0 + I - V}{\tau_m} + \sigma_m \sqrt{\tau_m} \xi(t)$$

## Time-resolvable interactions

Additionally we study two domains of interactions between populations; contemporaneous and delay-resolvable connections. These domains represent the relative timescales of measurement versus timescale of synaptic delay.

==DANGER doesnt work with pandoc==

correlation across positive and negative lags between two outputs

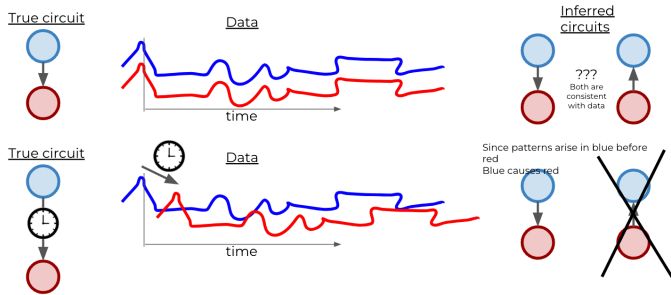
In the delay-resolvable domain, directionality of connections may be inferred even under passive observations by looking at temporal precedence - whether the past of one signal is more strongly correlated with future lags of another signal (*i.e. cross-correlation*). In the contemporaneous domain, network influences act within the time of a single sample<sup>8</sup> so this temporal precedence clue is lost (although directionality can still be inferred in the presence of intervention).

The following work is presented with the linear-Gaussian and contemporaneous domains as the default for simplicity and conciseness.

!!!! - talk about the extension to time-resolvable, spiking if it ends up being included

concept figures

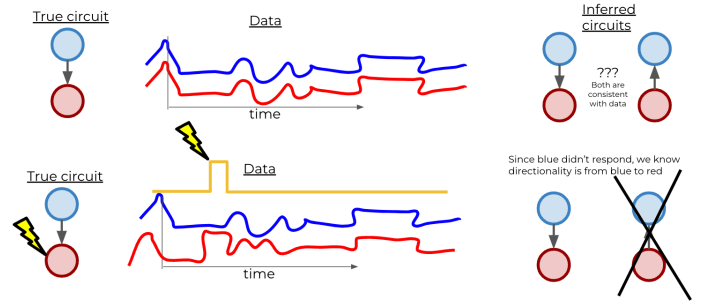
But **delayed connections** & observing temporal precedence can identify directionality from **passive observations alone**



<sup>7</sup>However, depending on overall firing rates and population sizes, this sparse spike-based transmission can be coarse-grained to a firing-rate-based model.

<sup>8</sup>the effective  $\Delta_{sample}$  would be broadened in the presence of jitter in connection delay, measurement noise, or temporal smoothing applied post-hoc, leading

## Open-loop stimulation helps disambiguate contemporaneous links



## Code implementation

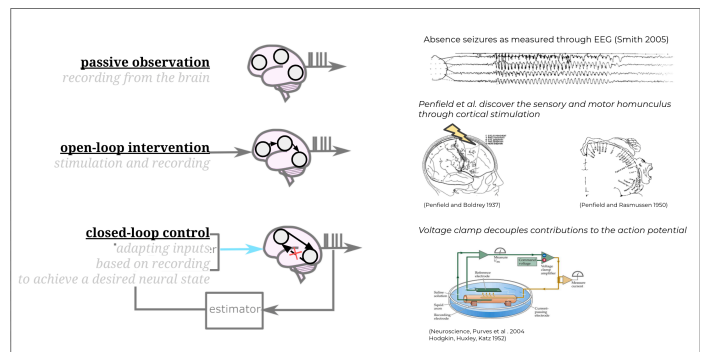
Software for data generation, analysis, and plotting is available at <https://github.com/awillats/clinc>. Both linear-gaussian and spiking networks are simulated with code built from the Brian2 spiking neural network simulator. This allows for highly modular code with easily interchanged neuron models and standardized output preprocessing and plotting. It was necessary to write an additional custom extension to Brian2 in order to capture delayed linear-gaussian interactions, available at [brian\\_delayed\\_gaussian](#). With this added functionality, it is possible to compare the equivalent network parameters only changing linear-gaussian versus spiking dynamics and inspect differences solely due to spiking.

!!!! - talk about parameter choices and ranges?

see `__network_parameters_table.md` for list of relevant parameters

## Implementing interventions

!!!! - 70% done !!!! - assumed: effect of interventions on theory already address



To study the effect of various interventions we simulated inputs to nodes in a network. In the **passive setting**, nodes receive additive drive from *private* Gaussian noise sources common to all neurons within a node, but independent across nodes. The variance of this noise is specified by  $\sigma_m \sqrt{\tau_m}$ .<sup>9</sup>

$$\frac{dV}{dt} = \frac{V_0 + I - V}{\tau_m} + \sigma_m \sqrt{\tau_m} \xi(t)$$

<sup>9</sup>need to triple check indexing w.r.t. nodes, neurons

To emulate **open-loop intervention** we simulated current injection from an external source. This is intended to represent experiments involving stimulation from microelectrodes or optogenetics (*albeit simplifying away any impact of actuator dynamics*). By default, open-loop intervention is specified as white noise sampled at each timestep from Gaussian distribution with mean and variance  $\mu_{intv.}$  and  $\sigma_{intv.}^2$ .<sup>10</sup>

$$I_{open-loop} \sim \mathcal{N}(\mu_{intv.}, \sigma_{intv.}^2)$$

Ignoring the effect of signal means in the linear-Gaussian setting:

$$X_k = f(\sigma_m^2, \sigma_{intv.}^2)$$

per-node indexing needs resolving here also

Ideal **closed-loop control** is able to overwrite the output of a node, setting it precisely to the specified target. making up notation as I go here, needs tightening up:

$$\begin{aligned} T &\sim \mathcal{N}(\mu_{intv.}, \sigma_{intv.}^2) \\ I_{closed-loop} &= f(X, T) \\ X_k | CL_k &\approx T \end{aligned}$$

Note that in this setting, the *output* of a node  $X_k$  under closed-loop control is identical to the target, therefore

$$X_k | CL_k = f(\sigma_{intv.}^2) \perp \sigma_m^2$$

In practice, near-ideal control is only possible with very fast measurement and computation relative to the network’s intrinsic dynamics, such as in the case of dynamic clamp<sup>11</sup>. To demonstrate a broader class of closed-loop interventions (such as those achievable with extracellular recording and stimulation), imperfect “partial” control is simulated by linearly interpolating the output of each node between the target  $T$  and the uncontrolled output based on a control effectiveness parameter  $\gamma$

$$X | CL_{k,\gamma} = \gamma T + (1 - \gamma) X$$

In the full discrete-time simulation, closed-loop interventions are instead simulated through a proportional-integral-derivative (PID) control policy with control efficacy determined functionally by the strength of controller gains  $K = \{k_P, k_I, k_D\}$  relative to the dynamics of the network.

$$I_{PID} = \text{PID}(X, T | K)$$

Another interesting intervention to study is **open-loop replay of a closed-loop stimulus**, *that is* taking a particular injected current  $I_{CL, prev}$  used to drive nodes to a target  $T_{prev}$  and adding it back to the network in a separate trial.

Because the instantiation of noise in the network will be different from trial to trial, this “replay” stimulus will no longer adapt sample-by-sample (therefore it should be considered

<sup>10</sup>need to resolve differences in implementation between contemporaneous and voltage simulation cases

<sup>11</sup>NEED dynamic clamp refs - [http://www.scholarpedia.org/article/Dynamic\\_clamp](http://www.scholarpedia.org/article/Dynamic_clamp)

open-loop) and the node’s output cannot be expected to match the target precisely, however the statistics of externally applied inputs will be the same. In effect, the comparison between closed-loop and open-loop replay conditions reveals the specific effect of feedback intervention while controlling for any confounds from input statistics.

## Extracting circuit estimates

!!!! - 10% done

*refer to methods overview figure*

While a broad range of techniques<sup>12</sup> exist for inferring functional relationships from observational data, (**for the majority of this work**) we choose to focus on simple bivariate correlation as a measure of dependence in the linear-Gaussian network. The impact of intervention on this metric is analytically tractable (*see methods1\_predicting\_correlation.md*), and can be thought of as a prototype<sup>13</sup> for more sophisticated measures of dependence such as time-lagged cross-correlations, bivariate and multivariate transfer entropy.

We implement a naive comparison strategy to estimate the circuit adjacency from empirical correlations; Thresholded empirical correlation matrices are compared to correlation matrices predicted from each circuit in a hypothesis set. Any hypothesized circuits which are predicted to have a similar correlation structure as is observed (i.e. corr. mats equal after thresholding) are marked as “plausible circuits.”<sup>14</sup> If only one circuit amongst the hypothesis set is a plausible match, this is considered to be the estimated circuit. The threshold for “binarizing” the empirical correlation matrix is treated as a hyperparameter to be swept at the time of analysis.<sup>15</sup>

## Information-theoretic measures of hypothesis ambiguity

!!!! - 10% done

*see \_steps\_of\_inference.md for entropy writeup*

## Results

!!!! - overall, 60% done

## Impact of intervention on estimation performance

<sup>12</sup>*inference techniques mentioned in the intro...*

<sup>13</sup>what does “prototype” mean here? something like MI and corr are equivalent in the linear-Gaussian case, ...

<sup>14</sup>TODO? formalize notation for this

<sup>15</sup>not sure how important this is. would prefer to set this threshold at some ad-hoc value since we’re sweeping other properties. But a more in-depth analysis could look at a receiver-operator curve with respect to this threshold



## Intervening provides (categorical) improvements in inference power beyond passive observation

!!!! - Application to demo set, entropy over hypotheses - 50% done

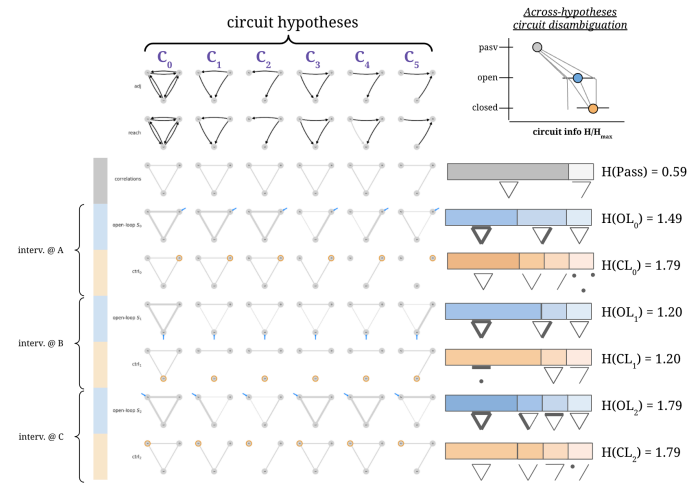
notes, see also

going to assume these have already been discussed

- predicting correlation
- measuring dependence
- markov equivalence

Methods: Procedure for choosing & applying intervention

Next, we apply (steps 1-3 of) this circuit search procedure to a collection of closely related hypotheses for 3 interacting nodes<sup>16</sup> to illustrate the impact of intervention. **most of the story in the figure caption for now**



**Figure DISAMBIG: Interventions narrow the set of hypotheses consistent with observed correlations** *source: google drawing* (A)

Directed adjacency matrices represent the true and hypothesized causal circuit structure (B) Directed reachability matrices represent the direct (*black*) and indirect (*grey*) influences in a network. Notably, different adjacency matrices can have equivalent reachability matrices making distinguishing between similar causal structures difficult, even with open-loop control. (C) Correlations between pairs of nodes. Under passive observation, the direction of influence is difficult to ascertain. In densely connected networks, many distinct ground-truth causal structures result in similar “all correlated with all” patterns providing little information about the true structure. (D-F) The impact of open-loop intervention at each of the nodes in the network is illustrated by modifications to the passive correlation pattern. Thick orange<sup>17</sup>

<sup>16</sup>nodes in such a graphical model may represent populations of neurons, distinct cell-types, different regions within the brain, or components of a latent variable represented in the brain.

<sup>17</sup>will change the color scheme for final figure. Likely using orange and blue to denote closed and open-loop interventions. Will also add

edges denote correlations which increase above their baseline value with high variance open-loop input. Thin blue<sup>18</sup> edges denote correlations which decrease, often as a result of increased connection-independent “noise” variance in one of the participating nodes. Grey edges are unaffected by intervention at that location. A given hypotheses set (A) will result in an “intervention-specific fingerprint”, that is a distribution of frequencies for observing patterns of modified correlations (*across a single row within D-F*). If this fingerprint contains many examples of the same pattern of correlation (such as B), many hypotheses correspond to the same observation, and that experiment contributes low information to distinguish between structures. A maximally informative intervention would produce a unique pattern of correlation for each member of the hypothesis set. :construction:caption too long

!!!! - Explain why closed-loop helps - link severing - 5% done

**Why does closed-loop control provide a categorical advantage?** Because it severs indirect links is this redundant with intro? needs to be backed here up by aggregate results? - this is especially relevant in recurrently connected networks where the reachability matrix becomes more dense. - more stuff is connected to other stuff, so there are more indirect connections, and the resulting correlations look more similar (more circuits in the equivalence class) - patterns of correlation become more specific with increasing intervention strength - more severed links → more unique adjacency-specific patterns of correlation

**Where you intervene<sup>19</sup>** strongly determines the inference power of your experiment. **secondary point:** having (binary) prediction helps capture this relationship

!!!! - Quantitative impact of closed-loop - 70% done  
### Stronger intervention shapes correlation, resulting in more data-efficient inference with less bias {#stronger-intervention-shapes-correlation-resulting-in-more-data-efficient-inference-with-less-bias }

!!!! - Explain why closed-loop helps - bidirectional variance control - 60% done [^dof]: need a more specific way of stating this. I mean degrees of freedom in the sense that mean and variance can be controlled independent of each other. And also, that the range of achievable correlation coefficients is wider for closed-loop than open-loop (where intrinsic variability constrains the minimum output variance)

While a primary advantage of closed-loop interventions for circuit inference is its ability to functionally lesion indirect connections, another, more nuanced (**quantitative**) ad-

in indication of severed edges

<sup>18</sup>will change the color scheme for final figure. Likely using orange and blue to denote closed and open-loop interventions. Will also add in indication of severed edges

<sup>19</sup>Figure VAR shows this pretty well, perhaps sink this section until after discussing categorical and quantitative?

vantage of closed-loop control lies in its capacity to bidirectionally control output variance. While the variance of an open-loop stimulus can be titrated to adjust the output variance at a node, in general, an open-loop stimulus cannot reduce this variance below its intrinsic<sup>20</sup> variability. That is, if the system is linear with gaussian noise,

!!!! todo - this is very closely related to 4.2 implementing interventions, description of impact of intervention on variance should perhaps be moved there... or the supplement?

$$\mathbb{V}_i(C|S = \text{open}, \sigma_S^2) \geq \mathbb{V}_i(C)$$

More specifically, if the open-loop stimulus is statistically independent from the intrinsic variability<sup>21</sup>

$$\mathbb{V}_i(C|S = \text{open}, \sigma_S^2) = \mathbb{V}_i(C) + \sigma_S^2$$

Applying closed-loop to a linear gaussian circuit:

$$\begin{aligned}\mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) &= \sigma_S^2 \\ \mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) &\perp \mathbb{V}_i(C)\end{aligned}$$

Firing rates couple mean and variance

In neural circuits, we're often interested in firing rates, which are non-negative. This particular output nonlinearity means that the linear gaussian assumptions do not hold, especially in the presence of strong inhibitory inputs. In this setting, firing rate variability is coupled to its mean rate; Under a homogeneous-rate Poisson assumption, mean firing rate and firing rate variability would be proportional. With inhibitory inputs, open-loop stimulus can drive firing rates low enough to reduce their variability. Here, feedback control still provides an advantage in being able to control the mean and variance of firing rates independently<sup>22</sup>

$$\begin{aligned}\mu_i^{\text{out}} &= f(\mu_i^{\text{in}}, \mathbb{V}_i^{\text{in}}) \\ \mathbb{V}_i^{\text{out}}(C) &= f(\mu_i^{\text{out}}, \mathbb{V}_i^{\text{in}})\end{aligned}$$

Notes on imperfect control

**Ideal control**

$$\mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) = \sigma_S^2$$

**Imperfect control** - intuitively feedback control is counteracting / subtracting disturbance due to unobserved sources, including intrinsic variability. We could summarize the effectiveness of closed-loop disturbance rejection with a scalar  $0 \leq \gamma \leq 1$

$$\mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) = \mathbb{V}_i(C) - \gamma \mathbb{V}_i(C) + \sigma_S^2 \mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) = (1 - \gamma) \mathbb{V}_i(C) + \sigma_S^2$$

!!!! - reference figvar to empirically show this bidirectional control of output variance?

<sup>20</sup>below the level set by added, independent/"private" sources

<sup>21</sup>notably, this is part of the definition of open-loop intervention

<sup>22</sup>practically, this requires very fast feedback to achieve fully independent control over mean and variance. In the case of firing rates, I suspect  $\mu \leq \alpha \mathbb{V}$ , so variances can be reduced, but for very low firing rates, there's still an upper limit on what the variance can be.

**Impact of intervention location and variance on pairwise correlations** related methods

!!!! TODO - again, feels very backgroundy / discussiony ... where to put this?

We have shown that closed-loop interventions provide more flexible control over output variance of nodes in a network, and that shared and independent sources of variance determine pairwise correlations between node outputs. Together, this suggests closed-loop interventions may allow us to shape the pattern of correlations with more degrees of freedom[~dof] [why do we want to?...]

One application of this increased flexibility is to increase correlations associated with pairs of directly correlated nodes, while decreasing spurious correlations associated with pairs of nodes without a direct connection (but perhaps are influenced by a common input, or are connected only indirectly). While "correlation does not imply causation," intervention may decrease the gap between the two.

Our hypothesis is that this shaping of pairwise correlations will result in reduced false positive edges in inferred circuits, "unblurring" the indirect associations that would otherwise confound circuit inference. However care must be taken, as this strategy relies on a hypothesis for the ground truth adjacency and may also result in a "confirmation bias" as new spurious correlations can be introduced through closed-loop intervention.

The impact of intervention on correlations can be summarized through the co-reachability  $\text{CoReach}(i, j|S_k)$ . A useful distillation of this mapping is to understand the sign of  $\frac{dR_{ij}}{dS_k}$ , that is whether increasing the variance of an intervention at node  $k$  increases or decreases the correlation between nodes  $i$  and  $j$

In a simulated network  $A \rightarrow B$  (fig. variance) we demonstrate predicted and empirical correlations between a pair of nodes as a function of intervention type, location, and variance. A few features are present which provide a general intuition for the impact of intervention location in larger circuits: First, interventions "upstream" of a true connection (lower left, fig. variance) tend to increase the connection-related variance, and therefore strengthen the observed correlations.

$$\text{Reach}(S_k \rightarrow i) \neq 0 \text{Reach}(i \rightarrow j) \neq 0 \frac{dR}{dS_k} > 0$$

Second, interventions affecting only the downstream node (lower right, fig. variance) of a true connection introduce variance which is independent of the connection  $A \rightarrow B$ , decreasing the observed correlation.

$$\text{Reach}(S_k \rightarrow j) = 0 \text{Reach}(S_k \rightarrow j) \neq 0 \frac{dR}{dS_k} < 0$$

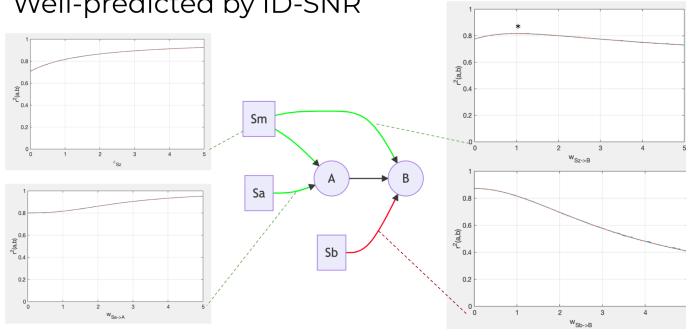
Third, interventions which reach both nodes will tend to increase the observed correlations (upper left, fig. variance), moreover this can be achieved even if no direct connection

$i \rightarrow j$  exists.

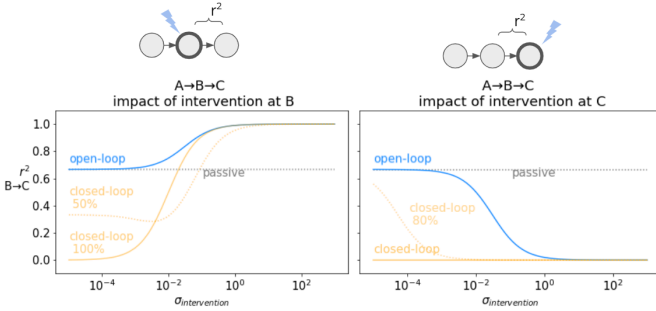
$$\text{Reach}(S_k \rightarrow i) \neq 0 \text{Reach}(S_k \rightarrow j) \neq 0 \text{Reach}(i \rightarrow j) = 0 \frac{dR}{dS_k} > 0$$

Notably, the impact of an intervention which is a “common cause” for both nodes depends on the relative weighted reachability between the source and each of the nodes. Correlations induced by a common cause are maximized when the input to each node is equal, that is  $\tilde{W}_{S_k \rightarrow i} \approx \tilde{W}_{S_k \rightarrow j}$  (upper right \* in fig. variance). If  $i \rightarrow j$  are connected  $\tilde{W}_{S_k \rightarrow i} \gg \tilde{W}_{S_k \rightarrow j}$  results in a variance-correlation relationship similar to the “upstream source” case (increasing source variance increases correlation  $\frac{dR}{dS_k} > 0$ ), while  $\tilde{W}_{S_k \rightarrow i} \ll \tilde{W}_{S_k \rightarrow j}$  results in a relationship similar to the “downstream source” case ( $\frac{dR}{dS_k} < 0$ )<sup>23</sup>

### Quantitative impact of parameters Well-predicted by ID-SNR



Closed-loop intervention enables **bidirectional control of correlation**  
Impact in a linear-gaussian chain, two intervention locations



(Final figure will be a mix of these two panels, caption will need updating) **Figure VAR: Location, variance, and type of intervention shape pairwise correlations (CENTER)** A two-node linear gaussian network is simulated with a connection from  $A \rightarrow B$ . Open-loop interventions (blue) consist of independent gaussian inputs with a range of variances  $\sigma_S^2$ . Closed-loop interventions (orange) consist of feedback control with an independent gaussian target with a range of variances. *Incomplete closed-loop interventions result in node outputs which are a mix of the control*

<sup>23</sup>not 100% sure this is true, the empirical results are really pointing to  $dR/dW < 0$  rather than  $dR/dS < 0$ . Also this should really be something like  $\frac{d|R|}{dS}$  or  $\frac{d r^2}{dS}$  since these effects decrease the *magnitude* of correlations. I.e. if  $\frac{d|R|}{dS} < 0$  increasing  $S$  might move  $r$  from  $-0.8$  to  $-0.2$ , i.e. decrease its magnitude not its value.

*target and network-driven activity.* Connections from sources to nodes are colored by their impact on correlations between A and B; green denotes  $dR/dS > 0$ , red denotes  $dR/dS < 0$ . **(lower left)** Intervention “upstream” of the connection  $A \rightarrow B$  increases the correlation  $r^2(A, B)$ . **(lower right)** Intervention at the terminal of the connection  $A \rightarrow B$  decreases the correlation  $r^2(A, B)$  by adding connection-independent noise. **(upper left)** Intervention with shared inputs to both nodes generally increases  $r^2(A, B)$ , (even without  $A \rightarrow B$ , see supplement). **(upper right)** The impact of shared interventions depends on relative weighted reachability  $\text{Reach}(S_k \rightarrow A)/\text{Reach}(S_k \rightarrow B)$ , with highest correlations when these terms are matched (see ) *Closed-loop interventions* (orange)\* generally result in larger changes in correlation across  $\sigma_S^2$  than the equivalent open-loop intervention. Closed-loop control at B effectively lesions the connection  $A \rightarrow B$ , resulting in near-zero correlation.<sup>24</sup>

additional notes:

- contextualize increasing correlation is sometimes good, sometimes bad!
- having (quantitative) prediction helps capture this relationship
- **(incidental) subfigure PREDICT: Comparing predicted and empirical correlation, identification performance**

The change in correlation as a function of changing intervention variance ( $\frac{d r_{ij}^2}{d S}$ ) can therefore be used as an additional indicator of presence/absence and directionality of the connection between A,B (see fig. disambig. D.))

Fig. variance also demonstrates the relative dynamic range of correlations achievable under passive, open- and closed-loop intervention. In the passive case, correlations are determined by intrinsic properties of the network  $\sigma_{base}^2$ . These properties have influence over the observed correlations in a way that can be difficult to separate from differences due to the ground-truth circuit. With open-loop intervention we can observe the impact of increasing variance at a particular node, but the dynamic range of achievable correlations is bounded by not being able to reduce variance below its baseline level. With closed-loop control, the bidirectional control of the output variance for a node means a much wider range of correlations can be achieved (blue v.s. orange in fig. variance), resulting in a more sensitive signal reflecting the ground-truth connectivity.

see also results1B\_data\_efficiency\_and\_bias.md

!!!! todo - comaprison signs in rows of DISAMBIG figure  
!!!! todo - merge from “box style” where entriire story is in caption, to having something in body of results text !!!!  
todo - write “explain why CL is better” section, ? exile

<sup>24</sup>compare especially to “Transfer Entropy as a Measure of Brain Connectivity”, “How Connectivity, Background Activity, and Synaptic Properties Shape the Cross-Correlation between Spike Trains” Figure 3.

it to discussion section? !!!! todo - connect DISAMBIG caption to quantitative variance explanation section !!!! todo - collapse figvar - do we need to make shared input point here? or is discussion fine? !!!! todo - dR/dS needs to mention R as  $r^2$  corr

Notes from matt

- [super minor] First part of fig DISAMBIG: subsections (A) through (C) work really well
- [super minor] in caption for (D-F): “modifications to the passive correlation pattern” is a bit confusing in the context of open-loop intervention
- [super minor] also in caption for (D-F): really like “intervention-specific fingerprint” terminology. The last sentence of the (D-F) caption really hits the message home, possible to emphasize that this is the take-home message earlier?
- [narrative/organization] fig DISAMBIG feels really example-y, more like a proof of concept than ‘results.’ The writing in Sec 5.1.1 also has this flavor, like it could be in a methods section. (The plot in the top right feels much more results-ey.) Not necessarily a bad thing, maybe just a consideration for thinking about article vs perspective flavor.
- [missing] Section 5.1.2.1: what are the definitions of  $S_k$ ,  $\text{CoReach}(i,j|S_k)$ , and  $R_{\{i\}}$ ?
- [narrative] Section 5.1.2.1: the narrative here really works for me, but it’s a little unclear whether this is more of a ‘result’ or a ‘recipe’ – the figures here also feel more example/proof-of-concept-ey, and the math here helps ground things in
- [missing] discussion of partial closed-loop control?

## Discussion

### limitations

The examples explored in this work simplify several key features that may have relevant contributions to circuit identification in practical experiments. [...]

### full observability

### results summary → summary of value closed-loop generally

Closed-loop control has the disadvantages of being more complex to implement and requires specialized real-time hardware and software, however it has been shown to have multifaceted usefulness in clinical and basic science applications. Here we focused on two advantages in particular; First, the capacity for functional lesioning which (reversibly) severs inputs to nodes and second, closed-loop control’s capacity to precisely shape variance across nodes. Both of these advantages facilitate opportunities for closed-loop intervention to reveal more circuit structure than passive observation or even open-loop experiments.

### summary of guidelines for experimentors

In studying the utility of various intervention for circuit inference we arrived at a few general guidelines which may assist experimental neuroscientists in designing the right intervention for the question at hand. First, more ambiguous hypotheses sets require “stronger” interventions to distinguish. Open-loop intervention may be sufficient to determine directionality of functional relationships, but as larger numbers of similar hypotheses [...] closed-loop intervention reduces the hypothesis set more efficiently. Second, we find that dense networks with strong reciprocal connections tend to result in many equivalent circuit hypotheses, but that well-placed closed-loop control can disrupt loops and simplify correlation structure to be more identifiable.<sup>25</sup> Recurrent loops are a common feature of neural circuit, and represent key opportunities for successful closed-loop intervention. The same is true for circuits with strong indirect correlations

### hidden confounds

### “funnel out”, future work → broad impact

### sequential experimental design

see *limitations\_future\_work.md*

## References

see *pandoc pandoc-citations*

## Supplement

“Adam-to-Do.” 2022.

Chicharro, Daniel, and Anders Ledberg. 2012. “When Two Become One: The Limits of Causality Analysis of Brain Dynamics.” Edited by Thomas Wennekers. *PLoS ONE* 7 (3): e32466. <https://doi.org/10.1371/journal.pone.0032466>.

Maathuis, Marloes H., and Preetam Nandy. 2016. “A Review of Some Recent Advances in Causal Inference.” In *Handbook of Big Data*, 387–408.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second. Cambridge University Press.

<sup>25</sup>this corroborates Ila Fiete’s paper on bias as a function of recurrent network strength