# Closed-Loop Identifiability in Neural Circuits

**Authors:** Adam Willats, Matt O'Shaughnessy

# Table of Contents



# Abstract

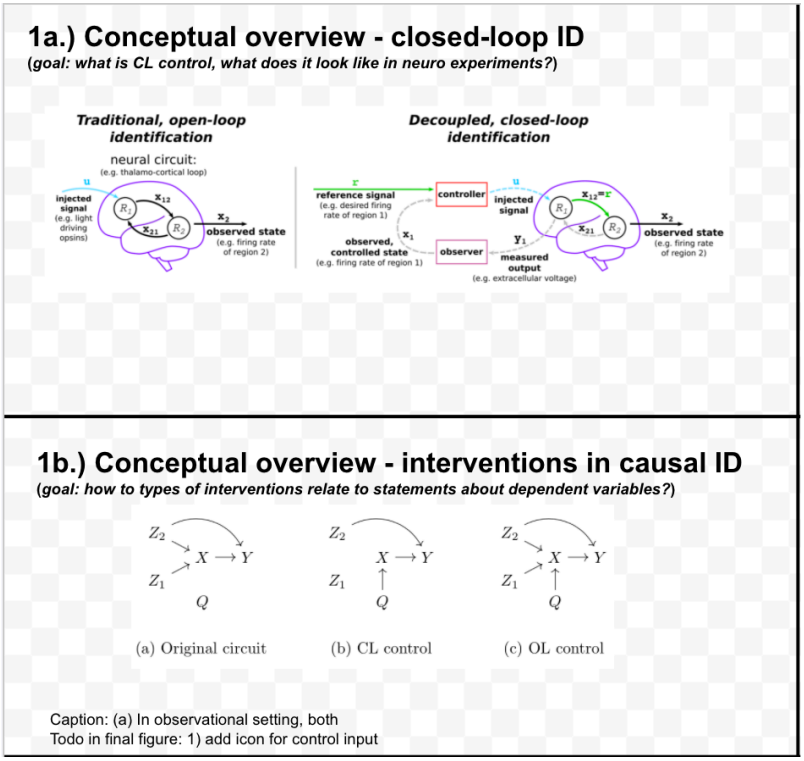
The necessity of intervention in inferring cause has long been understood in neuroscience. Recent work has highlighted the limitations of passive observation and single-site lesion studies in accurately recovering causal circuit structure. The advent of optogenetics has facilitated increasingly precise 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[1]. 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




**40% done:**

**our goal: estimating causal interactions in the brain**

Many hypotheses about neural circuits are best stated in terms of causal relationships: "changes made in to activity in this region of the brain will produce corresponding changes in that downstream region." Understanding these causal relationships is critical to developing effective therapeutic interventions, which require knowledge of how potential therapies will change 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 observational data, it is often impossible to determine whether observed patterns of activity are due to known and controlled inputs, or whether they are caused by recurrent activity, indirect relationships, or unseen "confounders." The

chemical and surgical lesion experiments that have historically been employed to remove the influence of possible confounds 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 \cite{chicharro2012when}.

In this paper we demonstrate when and how *closed-loop interventions* can reveal the causal structure governing neural circuits. 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, it is not yet fully understood when more complex intervention strategies can provide additional inferential power or how these interventions should be designed. To meet this need, we draw from tools used in causal inference \cite{pearl2009causality} \cite{maathuis2016review} \cite{chis2011structural}, which answer questions about what classes of models can be distinguished under a given set of input output experiments, and what experiments are necessary to determine internal connections uniquely.

We first propose a mathematical framework that describes how open- and closed-loop interventions impact the observable qualities of neural circuits. Using both simple controlled models and in silico models of neural circuits, 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 that can better uncover the connections which underly neural circuit function.

**how to infer causal interactions from time series?**

A number of measures have been proposed to quantify the strength of interaction between 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 \cite{wiener1956theory}. This requirement that *all* potentially causative variables be considered makes these notions of dependence susceptible to unobserved confounders (cite ?). Granger causality has traditionally been operationalized with vector autoregressive models \cite{granger1969investigating}. Drawbacks of Granger causality.

Our work initially focuses on measures of directional interaction that are based on cross-correlation. These metrics look at the correlation of time series collected from two nodes at various lags, taking a peak at a negative time lag as evidence for the existence of a potential causative relationship (more precise description). While cross-correlation-based measures are generally limited to detecting linear functional relationships between nodes, it is computationally inexpensive, making it a metric of choice for many real-world problems. More about correlation-based measures.

Other metrics quantify directional interaction stemming from nonlinear functional relationships (more precise description). Information-theoretic methods use information measures to assess the reduction in entropy knowledge of one variable provides about another, and is closely related to Granger causality in simple circuits \cite{barnett2009granger}. Transfer entropy ... \cite{bossomaier2016transfer}. Conditional transfer entropy ... (see bossomaier2016transfer sec 4.2.3) Definition, extension of intuition How conditional TE can address challenges ==Connection to causality (janzing2013quantifying, ay2008information, lizier2010differentiating) Estimation strategies (start with shorten2021estimating ?)

There are several important aspects to consider when comparing metrics for causal influence. bivariate vs multivariate approaches statistical testing [group effect/post-hoc tests; issues of multiple comparisons] note that we are leaning on IDTxl for this; no need to dive too deeply

GC in neuro [@Adam]

**todo - skim through these papers for methods/material to cite**

- reviews to read/cite:
    - broad background
        - \cite{TODO-runge2018causal} [in progress]
        - \cite{TODO-runge2019inferring}
        - \cite{TODO-peters2020causal}
    - specific to neuro
        - \cite{TODO-chicharro2012when}
        - \cite{TODO-dean2016dangers}
        - \cite{TODO-garofalo2009evaluation}
        - \cite{TODO-knox1981detection
        - \cite{TODO-salinas2001correlated}
        - \cite{TODO-wibral2014directed}
    - maybe...
        - \cite{TODO-lacasa2015network}
        - \cite{TODO-melssen1987detection}
- see `sketches_and_notation/background_why_control.md`

# Interventions in neuroscience & causal inference

**50% done:**

# Outline

- core idea is that "stronger" interventions lead to "higher inferential power"
    - may mean identifying circuits with less data
    - but may also mean distinguishing circuits which may have been "observationally equivalent" under weaker interventions
- **Highlight that the impact of interventions may generalize across any particular choice of inference algorithm**
- different effect of intervention types

# See also

- notation1_circuits.md
- notation2_do_calculus.md
- notation3_pearl.md

# Draft

A rich theoretical literature has confirmed the central role of interventions in inferring causal structure from data

\cite{pearl2009causality, eberhardt2007interventions}. Consistent with intuition from neuroscience literature, data in which some variables are experimentally intervened on is typically much more powerful than observational data alone. For example, observational data of two correlated variables $x$ and $y$ does not allow a scientist to determine whether $x$ is driving $y$, $y$ is driving $x$, or if the two variables are being independently driven by a hidden confounder. Experimentally manipulating $x$ and observing the output of $y$, however, allows the scientist to begin to establish which potential causal interaction pattern is at work.

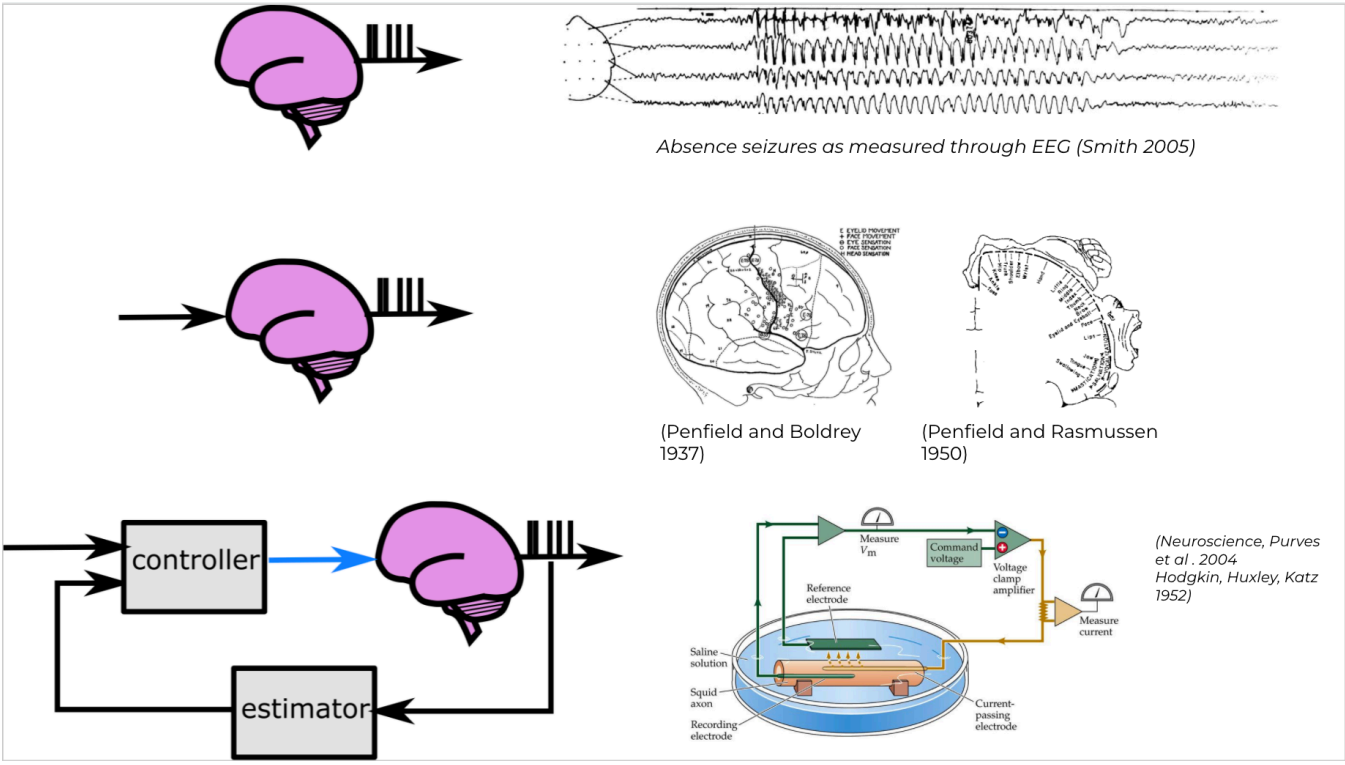



**Fig: Examples of the role of interventions in discoveries in neuroscience** (A) Identifying when a patient is having a seizure, from **passive recordings** alone (B) through **systematic open-loop stimulation experiments,** Penfield was able to uncover the spatial organization of how senses and movement are mapped in the cortex [2] (C) **Feedback control** allows us to specify activity in the brain in terms of outputs. Allows us to reject disturbances, respond to changes

Certain interventions can provide much more inferential power than others.[3] Interventions on some portions of a system may allow more information about the system's causal structure than interventions in other areas. Interventions are also more valuable when they more precisely change the system: "perfect" interventions that set the behavior of part of the system exactly to a desired state provide more information than "soft" interventions that only partially manipulate a part of the system.

In real-world neuroscience settings, experimenters are faced with deciding between interventions that differ in both of these regards. For example, stimulation can often only be applied to certain regions of the brain [4]. And while experimenters may be able to exactly manipulate activity in some parts of the brain using closed-loop control (akin to a "perfect" intervention), in other locations experimenters may only be able to apply open-loop control that perturbs a part of the system but can not manipulate its activity exactly to a desired state (a "soft" intervention).

Although theoretical guarantees and algorithms designed to choose among these interventions are often designed for simple models with strong assumptions on properties such as the types of functional relationships that exist in circuits, the visibility

and structure of confounding relationships, and noise statistics, they provide guidance that can that can help practitioners design experiments that provide as much scientific insight as possible[5]
\cite{ghassami2018budgeted,yang2018characterizing}. Importantly, the necessity and inferential power of interventions is often \emph{algorithm-independent}, in the sense that there exist interventions that reveal causal structure that would be impossible for \emph{any} algorithm to infer from observational data alone \cite{shanmugam2015learning}.

In this paper, we take a theoretically- and experimentally-motivated approach to analyzing the ability of neurally-plausible open- and closed-loop interventions to provide information about the causal structure of neural circuits.

# Representations & reachability

> **60% done:**

↪ graph for shared v.s. private sources
graph TD eA-->A u(u)-->A u-->C A-->B C-->B eB-->B eC-->C
Different mathematical representations of circuits can elucidate different connectivity properties. For example, consider the circuit $A \to B \leftarrow C$. This circuit can be modeled by the dynamical system

$$\begin{cases} \dot{x}_A &= f_A(e_A) \\ \dot{x}_B &= f_B(x_A, x_C, e_B) \\ \dot{x}_C &= f_C(e_C), \end{cases}$$

where $e_A$, $e_B$, and $e_C$ represent exogenous inputs that are inputs from other variables and each other[6].

When the system is linear we can use matrix notation to denote the impact of each node on the others. Denote the $p \times n$ matrix of data samples by $X$ and the $p \times n$ matrix of exogenous input values by $E$. We can then write[7]

$$X = XW + E,$$

Topologically sorted implementation:

$$X^{-} := E \tag{1}$$
$$X := X^{-}W + E \tag{2}$$

> **TODO Adam, write out the dynamical system version of this**

where $W$ represents the *adjacency matrix*

$$W = \begin{bmatrix} w_{AA} & w_{AB} & w_{AC} \\ w_{BA} & w_{BB} & w_{BC} \\ w_{CA} & w_{CB} & w_{CC} \end{bmatrix}.$$

In the circuit $A \to B \leftarrow C$, we would have $w_{AB} \neq 0$ and $w_{CB} \neq 0$.

The adjacency matrix captures directional first-order connections in the circuit: $w_{ij}$, for example, describes how activity in $x_j$ changes in response to activity in $x_i$.

Our goal is to reason about the relationship between underlying causal structure (which we want to understand) and the correlation or information shared by pairs of nodes in the circuit (which we can observe). Quantities based on the adjacency matrix and weighted reachability matrix bridge this gap, connecting the causal structure of a circuit to the correlation structure its nodes will produce.

The directional $k^{\text{th}}$-order connections in the circuit are similarly described by the matrix $W^k$, so the *weighted reachability matrix*

$$W = \sum_{k=0}^{\infty} W^k$$

describes the total impact --- through both first-order (direct) connections and higher-order (indirect) connections --- of each node on the others. Whether node $j$ is "reachable" (Skiena 2011) from node $i$ by a direct or indirect connection is thus indicated by $W_{ij} \neq 0$, with the magnitude of $W_{ij}$ indicating sensitive node $j$ is to a change in node $i$.

This notion of reachability, encoded by the pattern of nonzero entries in $W$, allows us to determine when two nodes will be correlated (or more generally, contain information about each other). Moreover, as we will describe in Sections [REF] and [REF], quantities derived from these representations can also be used to describe the impact of open- and closed-loop interventions on circuit behavior, allowing us to quantitatively explore the impact of these interventions on the identifiability of circuits.

`[Matt to Adam --- I like the idea of an example here, but the details will likely need to change once the neighboring int`

Consider, for example, the hypotheses for cortical gain control in open-loop (Figure BACKGROUND>REPRESENTATION/REACH-1, left column). In both circuit 2a and 2b, PV cells are reachable from the Som cell node ($W_{PV \to Som} \neq 0$), since Som activity can influence PV activity indirectly through the Pyr node. These circuits are therefore difficult to distinguish under open-loop intervention.

If the reachability of two circuits are unequal for a given intervention, differences in correlation between observed regions will be sufficient to distinguish between the two hypotheses. Looking at these same circuits under closed-loop control of the pyramidal population (Figure BACKGROUND>REPRESENTATION/REACH-1, right column), dashed lines reveal that there is no longer an indirect functional connection from Som to PV cells. As such, in circuit 2a, PV cells are no longer reachable from the Som population, whereas they are reachable under circuit 2b. This difference in reachability corresponds to the difference in correlational structure that allows us to distinguish these two hypotheses under closed-loop control.

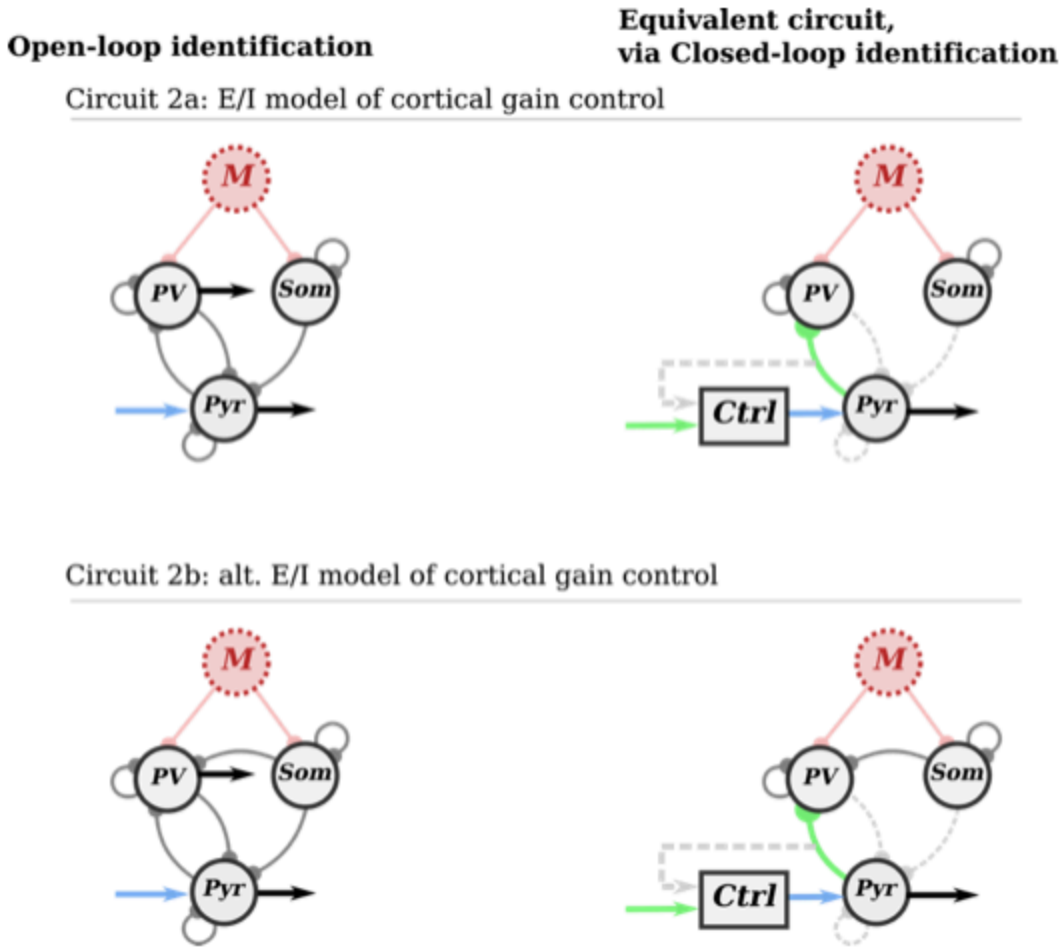



**Figure BACKGROUND>REPRESENTATION/REACH-1: Closed-loop control allows for two circuit hypotheses to be distinguished.** Two hypothesized circuits for the relationships between pyramidal (Pyr, excitatory), parvalbumin-positive (PV, inhibitory), and somatostain-expressing (Som, inhibitory) cells are shown 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.

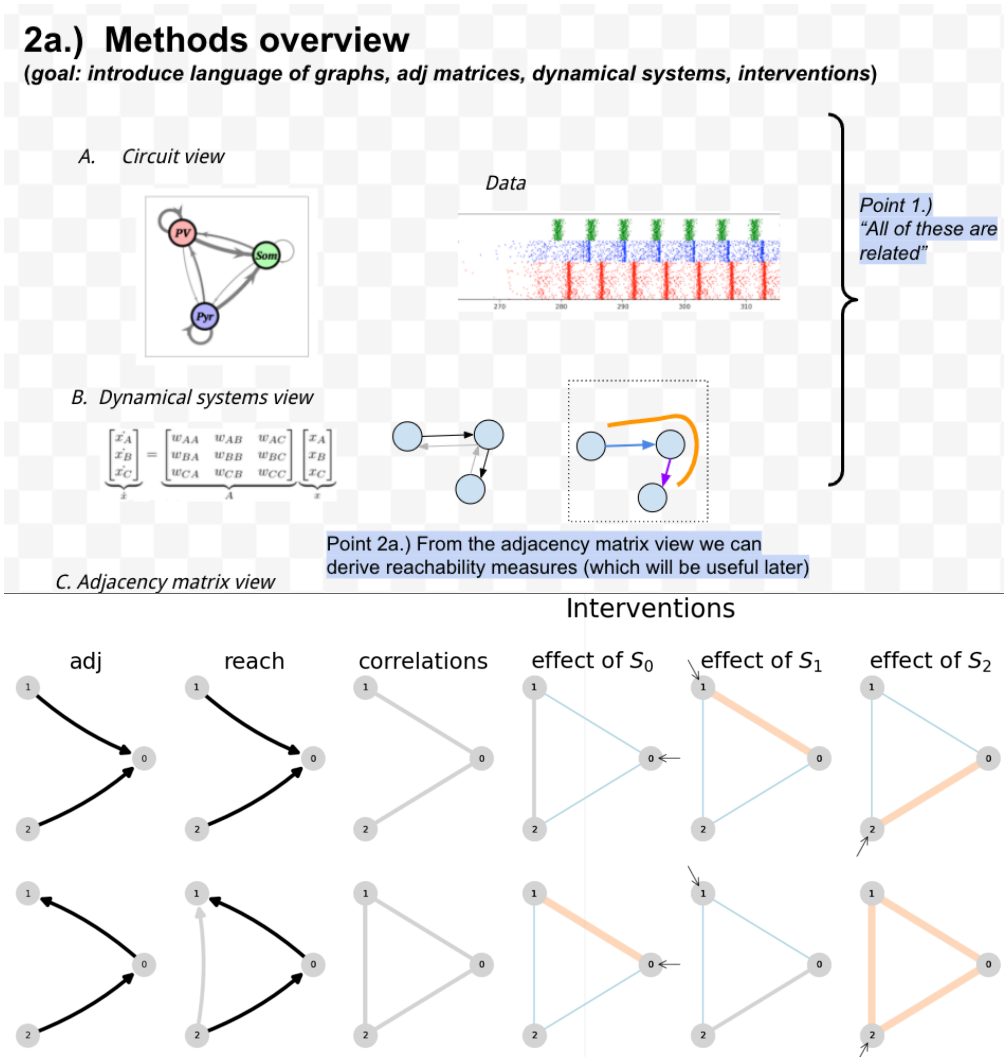



**Figure DEMO: Applying CLINC to distinguish a pair of circuits**

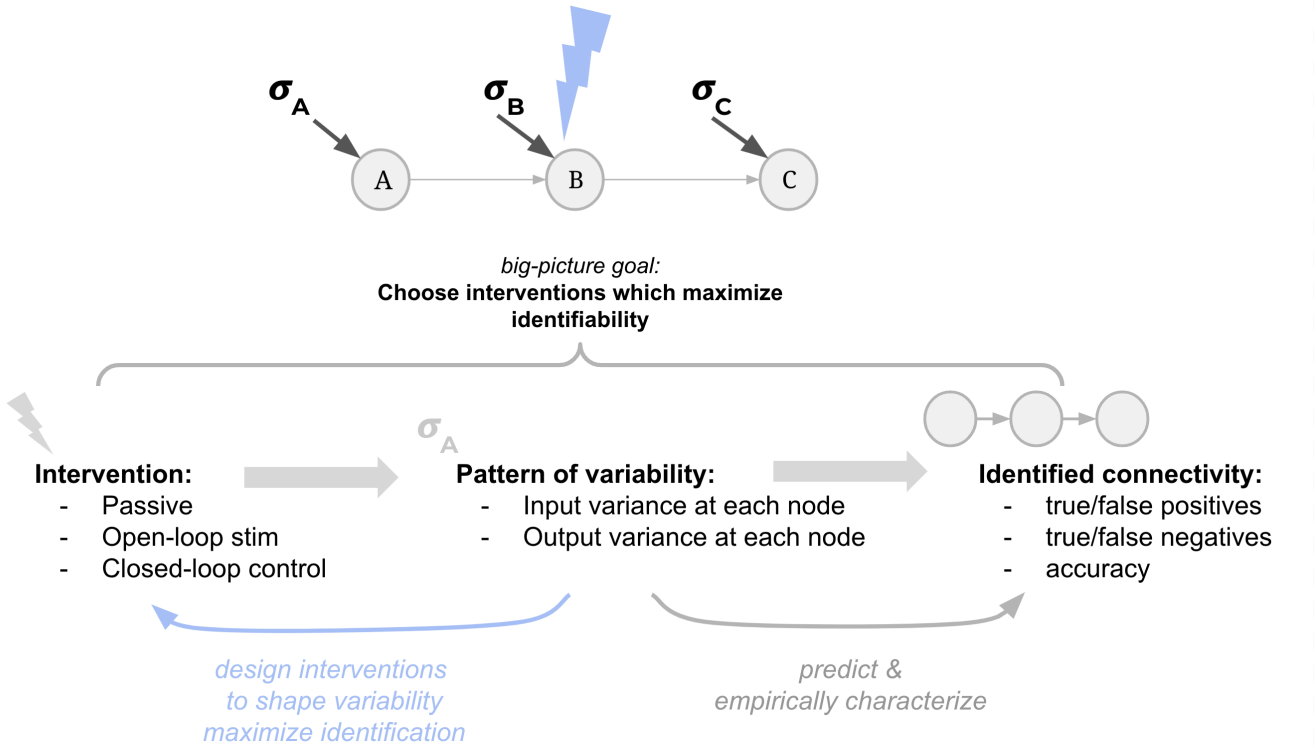
---

# Theory / Prediction

> *(OVERVIEW)*

## Computing reachability (theory)

## Predicting correlation structure (theory)




# Predicting network correlations

## Building blocks

Let $s \in \mathbb{R}^p$ denote exogenous inputs to each of the $p$ nodes, and $W \in \mathbb{R}^{p \times p}$ denote the matrix of connection strengths such that

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

We'll assume for now that all functional relationships between nodes are linear, and all exogenous noise is iid gaussian.[^lingauss]

For intuition, note that:

- $(s)_j$ denotes variance in node $j$ due to length-0 connections,
- $(Ws)_j$ denotes variance in node $j$ due to length-1 connections
- $(W^2 s)_j$ denotes variance in node $j$ due to length-2 connections
- ...
- $(\sum_{i=1}^{p} W^{i-1} s)_j$ denotes the total variance in node $j$.

## Derivation of expression for (co)variances

In a linear/gaussian network, we have $x = Wx + e$, where $x \in \mathbb{R}^p$ is the vector of values taken by the $p$ nodes of the circuit and $e \sim \mathcal{N}(0, \text{diag}(s))$. Rearranging this expression yields $X = (I - W)^{-1}s$.

Defining $X \in \mathbb{R}^{p \times n}$ and $E \in \mathbb{R}^{p \times n}$ as the matrix of $n$ observations of $x$ and $s$, we can write

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

It is a fact that $(I - A)^{-1} = \sum_{n=0}^{\infty} A^n$ when $|\lambda_i(A)| < 1$ for all eigenvalues $\lambda_i$ of $A$. We'll make use of the matrix $W = \sum_{k=0}^{p-1} W^k$, which intuitively denotes the "effective" or "total" connection strengths between every pair of nodes in the circuit, including both direct and indirect links. That is, $W_{ij}$ tells us how much variance at node $i$ would result from injecting a unit of variance at node $j$.[8]

To simplify a bit, we can equivalently write

$$\Sigma_{ij} = \sum_{k=1}^{p} W_{ik}W_{jk}s_k.$$

## Expression for $r(i, j)$ under passive observation

Using the expression for $\Sigma$ above, we have

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

# Impact of control

## Open-loop control

The application of open-loop control on node $c$ can be modeled as:

1. Add an arbitrary amount of variance to $s_c$: $s_c \leftarrow s_c + s_c^{(OL)}$.

## Closed-loop control

The application of closed-loop control on node $c$ can be modeled as:

1. Sever the inputs to node $c$ by setting $W_{c,:} = 0$, then
2. Set the exogenous noise of node $c$ by setting $s_c$ to any arbitrary value. Because $c$'s inputs have been severed, this exogenous noise will be node $c$'s output variance.

$\sim$ $\sim$

Note that step 1 above will result in $W_{i,:} = 0$ except for $W_{i,i} = 1$.

> **TODO for Adam, I think there's more interprettable stuff that can be said about the impact of closed-loop here. Largely based on intuition from the binary version**

### Impact of CL control on $r(i, j)$

We might be interested in $\Delta_c^{(CL)} r(i, j)$, defined as "the amount that $r(i, j)$ increases when we place closed-loop control on node $c$." Unfortunately, writing out a general expression for this gets ugly fairly quickly.

---

# Simulation

## Network simulations (simulation)

## Implementing interventions (simulation)

## Extracting circuit estimates (empirical)

## Information-theoretic measures of hypothesis ambiguity

---

# Results

## Interaction of intervention on circuit estimation

going to assume these have already been discussed

- predicting correlation
- measuring dependence
- markov equivalence

---

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

Methods: Procedure for choosing & applying intervention

---

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

Next, we apply (steps 1-3 of) this circuit search procedure to a collection of closely related hypotheses for 3 interacting nodes[9] to illustrate the impact of intervention. 🚧 `most of the story in the figure caption for now` 🚧

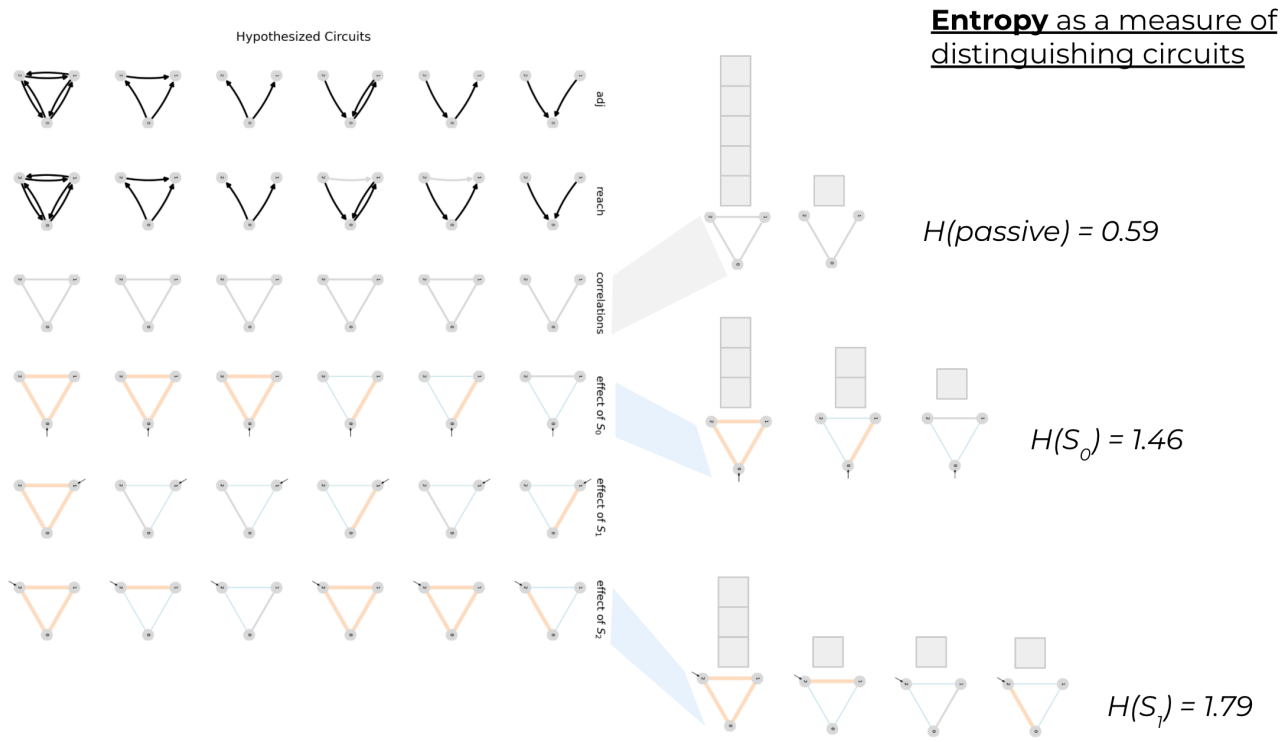



**Figure DISAMBIG: Interventions narrow the set of hypotheses consistent with observed correlations**
**(A)** Directed adjacency matrices represent the true and hypothesized causal circuit structure
**(B)** Directed reachability matrices represent the direct *(black)* and indirect *(grey)* influences 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[10] edges denote correlations which increase above their baseline value with

high variance open-loop input. Thin blue[10:1] 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.
🚧 `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**[11] 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

**Explain why closed-loop helps - bidirectional variance control - 60% done**

While a primary advantage of closed-loop interventions for circuit inference is its ability to functionally lesion indirect connections, another, more nuanced `(quantitative)` advantage 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 instrinsic[12] variability. That is, if the system is linear with gaussian noise,

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

More specifically, if the open-loop stimulus is statistically independent from the intrinsic variability[13]

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

Applying closed-loop to a linear gaussian circuit:

$$\mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) = \sigma_S^2 \tag{3}$$
$$\mathbb{V}_i(C|S = \text{closed}, \sigma_S^2) \perp \mathbb{V}_i(C) \tag{4}$$

↪ 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 homoeneous-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[14]

$$\mu_i^{out} = f(\mu_i^{in}, \mathbb{V}_i^{in}) \tag{5}$$
$$\mathbb{V}_i^{out}(C) = f(\mu_i^{out}, \mathbb{V}_i^{in}) \tag{6}$$

↪ 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 empricially show this bidirectional control of output variance?**

## Impact of intervention location and variance on pariwise correlations

related methods

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[15]

`[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). This manipulation may bring the observed pattern of correlations

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 $\mathrm{CoReach}(i, j|S_k)$. 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→B (fig. variance) we demonstrate predicted and emprirical 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.

$$\begin{gathered}\mathrm{Reach}(S_k \to i) \neq 0 \\ \mathrm{Reach}(i \to j) \neq 0 \\ \frac{dR}{dS_k} > 0\end{gathered}$$

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

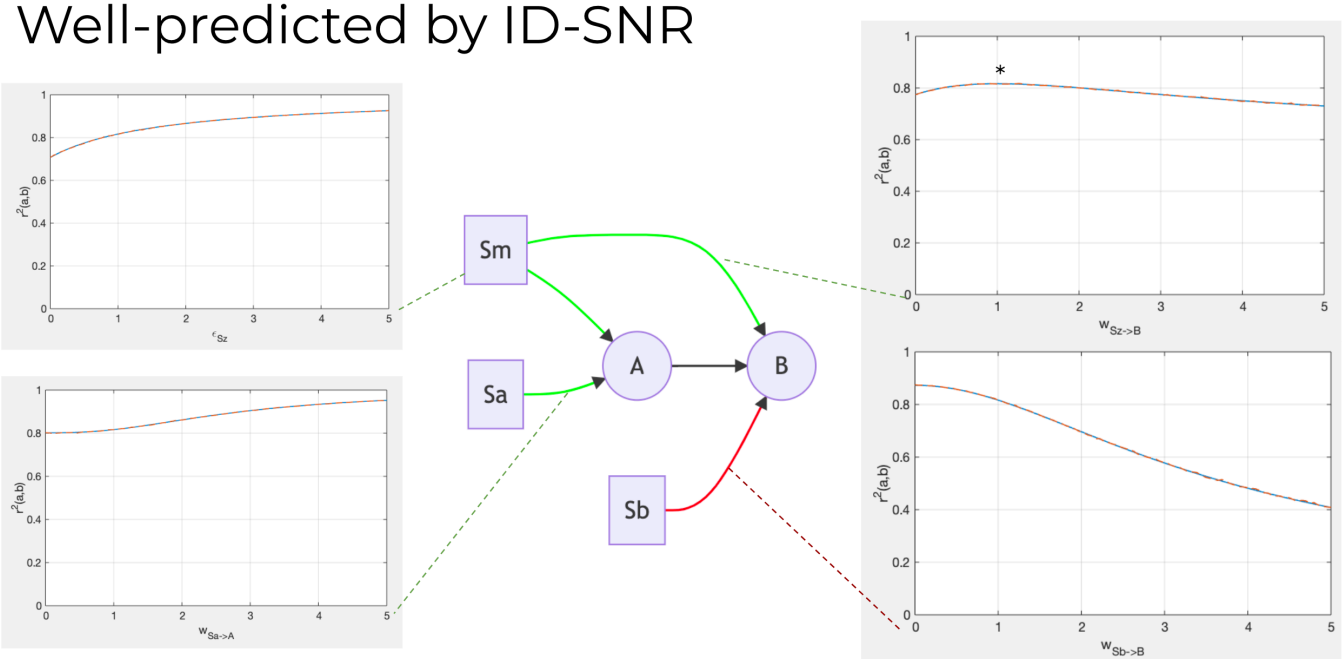
$$\begin{gathered}\mathrm{Reach}(S_k \to j) = 0 \\ \mathrm{Reach}(S_k \to j) \neq 0 \\ \frac{dR}{dS_k} < 0\end{gathered}$$

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 $i \to j$ exists.

$$\begin{gathered}\mathrm{Reach}(S_k \to i) \neq 0 \\ \mathrm{Reach}(S_k \to j) \neq 0 \\ \mathrm{Reach}(i \to j) = 0 \\ \frac{dR}{dS_k} > 0\end{gathered}$$

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 $W_{S_k \to i} \approx W_{S_k \to j}$ (upper right * in fig. variance). If i→j are connected $W_{S_k \to i} \gg W_{S_k \to j}$ results in an variance-correlation relationship similar to the "upstream source" case (increasing source variance increases correlation $\frac{dR}{dS_k} > 0$),
while $W_{S_k \to i} \ll W_{S_k \to j}$ results in a relationship similar to the "downstream source" case ($\frac{dR}{dS_k} < 0$)[16]

# **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*

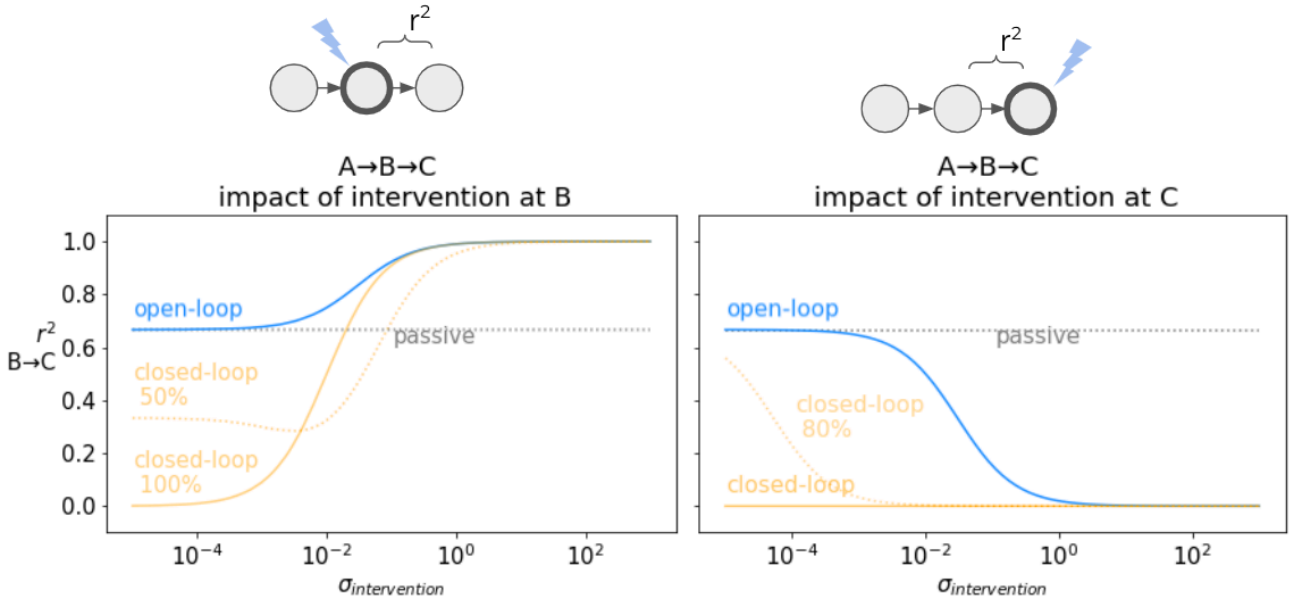



**Figure VAR: Location, variance, and type of intervention shape pairwise correlations**

**(CENTER)** A two-node linear gaussian network is simulated with a connection from A→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 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→B increases the correlation $r^2(A, B)$.
**(lower right)** Intervention at the terminal of the connection A→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→B, see supplement)*.
**(upper right)** The impact of shared interventions depends on relative weighted reachability $\mathrm{Reach}(S_k \to A)/\mathrm{Reach}(S_k \to B)$, 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→B, resulting in near-zero correlation. [17]

↪ 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**

🚧

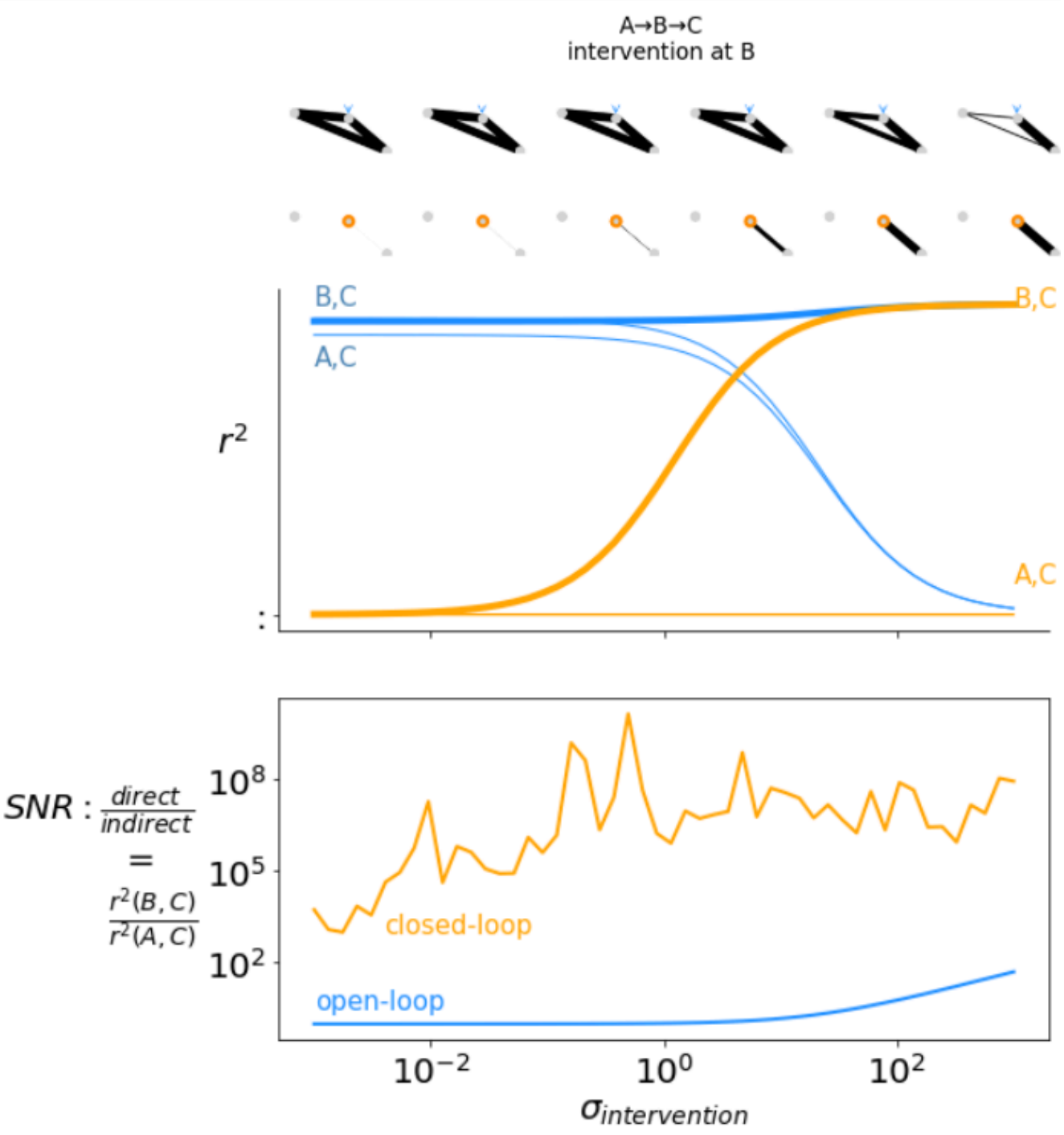
The change in correlation as a function of changing intervention variance ($\frac{dr_{ij}^2}{dS}$) can therefore be used as an additional indicator of presence/absence and directionality of the connection between A,B *(see fig. disambig. D.)*
🚧

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 instrinsic 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.

**Explain why closed-loop helps - more data efficient - 10% done**

- less data required to get to threshold level of accuracy (more data-efficient)
  - likely comes from improved "SNR" which can be thought of as a derived property of the per-edge correlations

🚧


A→B→C
intervention at B
B,C
A,C
B,C
A,C
$r^2$
$SNR : \frac{direct}{indirect} = \frac{r^2(B,C)}{r^2(A,C)}$
$10^8$
$10^5$
$10^2$
closed-loop
open-loop
$10^{-2}$
$10^0$
$10^2$
$\sigma_{intervention}$

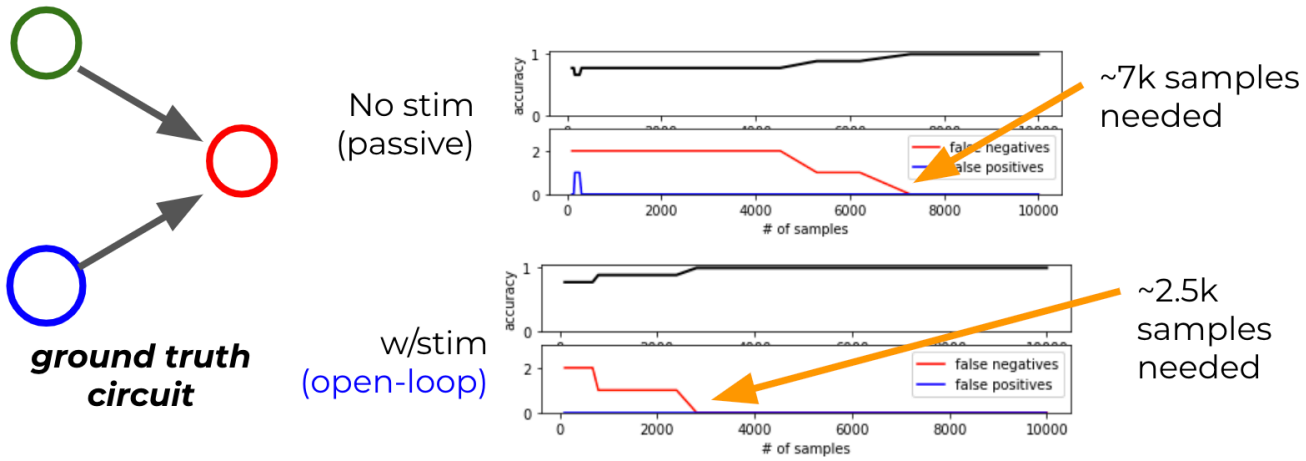



figure sketches

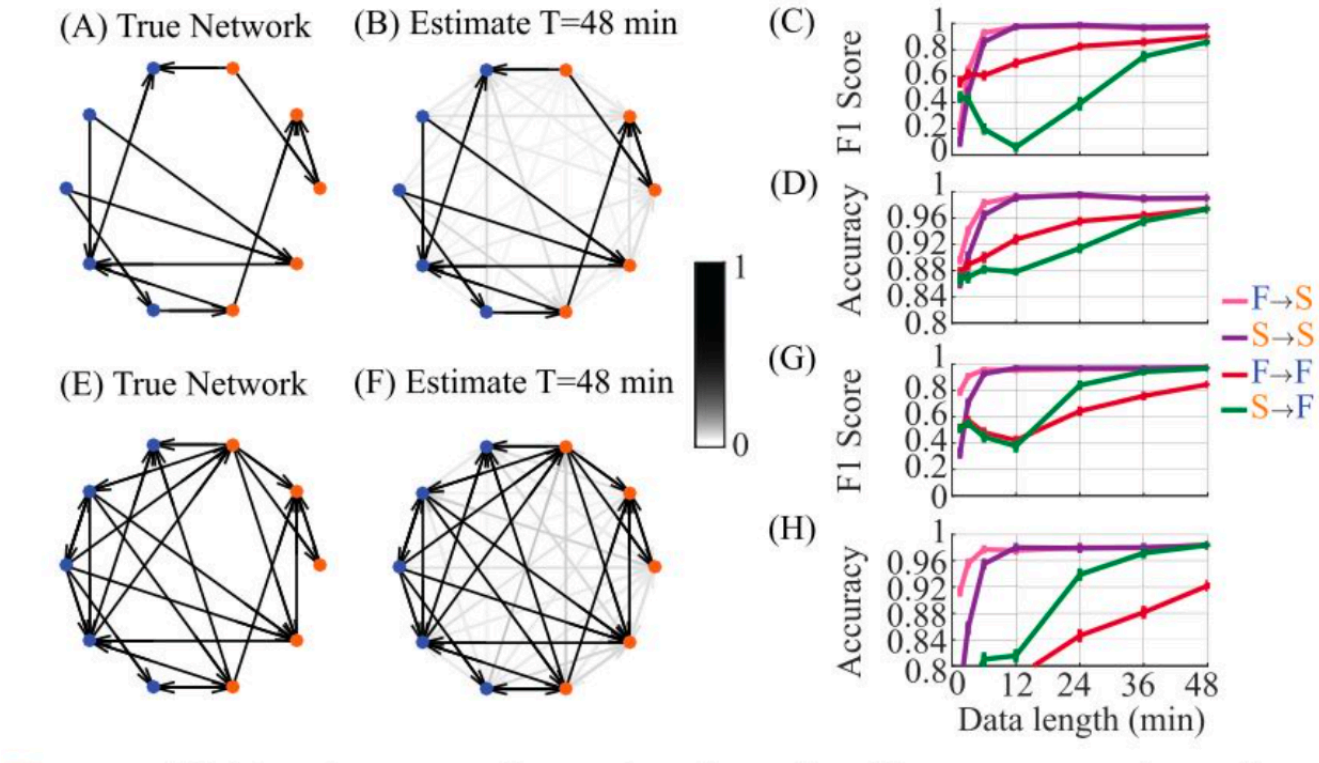



Figure DATA: Analysis of simulated circuits suggest stronger intervention facilitates identification with less data [18]

**Explain why closed-loop helps - less bias - 5% done**

- higher infinite-data accuracy (i.e. less bias)
    - lower bias likely comes from the categorical advantages above
- breakdown false positives, false negatives

🚧

## Interaction of intervention & circuit structure

**Figure MOTIF: Interaction of network structure and intervention location on identifiability**

---

# Discussion

*see limitations_future_work.md*

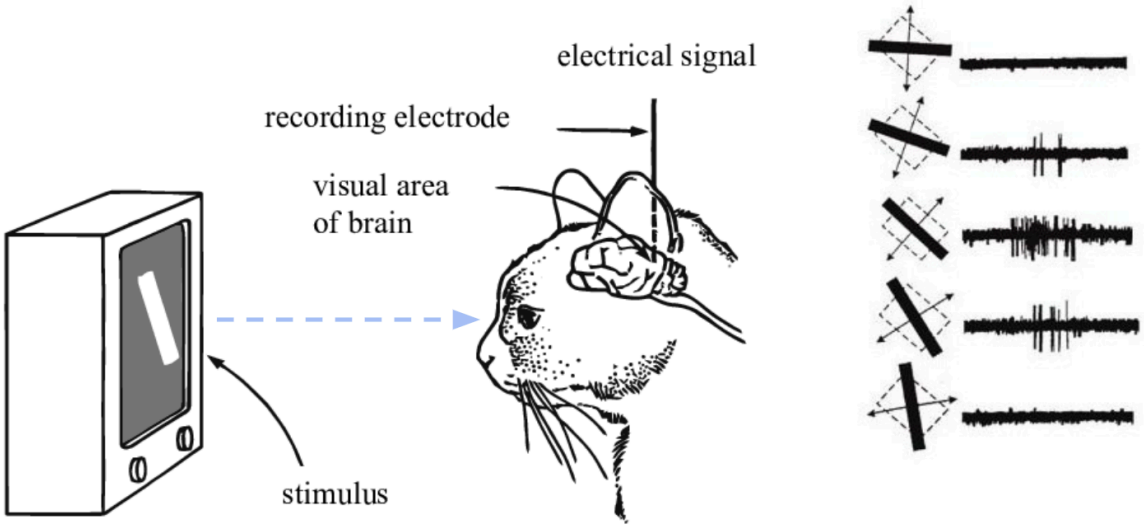
# References

*see pandoc pandoc-citations*

# Supplement

---

1. may end up discussing quantitative advantages such as bidirectional variance (and correlation) control ↩
2. Another great example of open-loop mapping: Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology 148(3), 574–591 (1959)




↩
3. probably needs to get more specific sooner, @Adam can fill in ↩
4. @Adam - make this more precise. talk about spatial, temporal degrees of freedom ↩
5. @Matt this needs breaking down ↩
6. the most important property of $e$ for the math to work, i believe, is that they're random variables independent of each other. This is not true in general if E is capturing input from common sources, other nodes in the

network. I think to solve this, we'll need to have an endogenous independent noise term and an externally applied (potentially common) stimulus term. ↩

7. have to be careful with this. this almost looks like a dynamical system, but isn't. In simulation we're doing something like an SCM, where the circuit is sorted topologically then computed sequentially. And then I'm ↩
8. We can use $p-1$ as an upper limit on the sum $W = \sum_{k=0}^{p-1} W^k$ when there are no recurrent connections. Later we can characterize what type of recurrent connections are ok. ↩
9. 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. ↩
10. 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 ↩ ↩
11. Figure VAR shows this pretty well, perhaps sink this section until after discussing categorical and quantitative? ↩
12. below the level set by added, independent/"private" sources ↩
13. notably, this is part of the definition of open-loop intervention ↩
14. 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. ↩
15. 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 instrinsic variability constrains the minimum output variance) ↩
16. 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 $\frac{dr^2}{dS}$ 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. ↩
17. 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. ↩
18. "Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model", "Evaluation of the Performance of Information Theory- Based Methods and Cross-Correlation to Estimate the Functional Connectivity in Cortical Networks" ↩