# 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

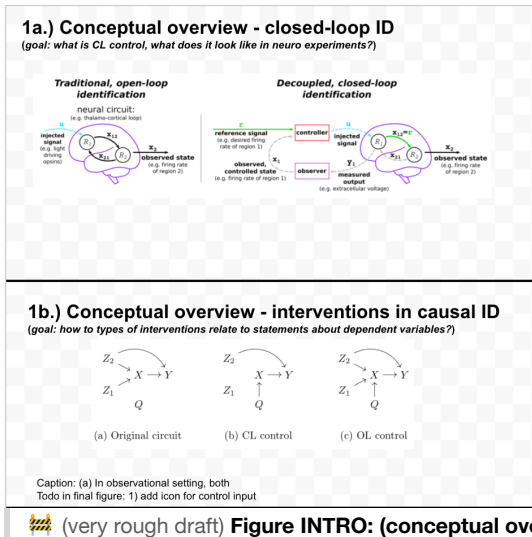## Estimating causal interactions in the brain

**1a.) Conceptual overview - closed-loop ID**
*(goal: what is CL control, what does it look like in neuro experiments?)*

**1b.) Conceptual overview - interventions in causal ID**
*(goal: how to types of interventions relate to statements about dependent variables?)*

(a) Original circuit    (b) CL control    (c) OL control

Caption: (a) In observational setting, both
Todo in final figure: 1) add icon for control input

🚧 (very rough draft) **Figure INTRO: (conceptual overview of interacting regions, intervention, DAGs etc.)**

✏ **40% done:**

**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.

**Inferring 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}, a concept that has traditionally been operationalized with vector autoregressive models \cite{granger1969investigating}. The requirement that *all* potentially causative variables be considered makes these notions of dependence susceptible to unobserved confounders \cite{runge2018causal}.

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, treating peaks at negative time lags as evidence for potential causative relationships. Such peaks could indicate the presence of a direct causal relationship -- but they could also stem from indirect causal links or hidden confounders. Unlike some multivariate methods that "control" for the effects of potential confounders, as a bivariate method it is necessary to consider patterns of correlation between many pairs of nodes in order to differentiate between direct, indirect, and confounding relationships \cite{dean2016dangers??-if no better source}. Cross-correlation-based measures are generally limited to detecting linear functional relationships between nodes, but their computational feasibility makes them a metric of choice in experimental neuroscience work \cite{knox1981detection} \cite{salinas2001correlated}.

Other metrics quantify 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 \cite{schreiber2000measuring} \cite{barnett2009granger}. The *transfer entropy* $T_{X \to Y}(t) = I(Y_t : X_{<t} \mid Y_{<t})$ 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 $Y_{<t}$) \cite{bossomaier2016transfer}. Using transfer entropy as a measure of causal interaction requires accounting for potential confounding variables; the *conditional transfer entropy* $T_{X \to Y|Z}(t) = I(Y_t : X_{<t} \mid Y_{<t}, Z_{<t})$ conditions on the past of other variables to account for their confounding influence \cite[Sec.~4.2.3]{bossomaier2016transfer}. Conditional transfer entropy can thus be interpreted as the amount of information present in $Y$ that is not present in either the past of $X$, the past of $Y$, or the past of other variables $Z$.

Information-theoretic and transfer-entropy-based methods used to quantify the strength of causal interactions typically require knowledge of the ground truth causal relationships that exist \cite{janzing2013quantifying} or the ability to perturb the system \cite{ay2008information} \cite{lizier2010differentiating}. In practice, these quantities are often interpreted as "information transfer," and a variety of estimation strategies and methods to automatically select variables and time lags to condition are used (e.g., \cite{shorten2021estimating}). Multivariate conditional transfer entropy approaches using various variable selection schemes can differentiate between direct interactions, indirect interactions, and common causes. The results from these methods can often differ based on 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 IDTxl in our results: In our empirical results using transfer-entropy-based notions of directional in Despite their mathematical differences, however, previous work has found that cross-correlation-based metrics and information-based metrics tend to produce qualitatively similar results \cite{garofalo2009evaluation} \cite{maybe: ito2011extending (shows CC generally underperforms their TE method, but mostly similar??)}.

▶ ↪reviews to read/cite: (todo)
*see also why_control.md*

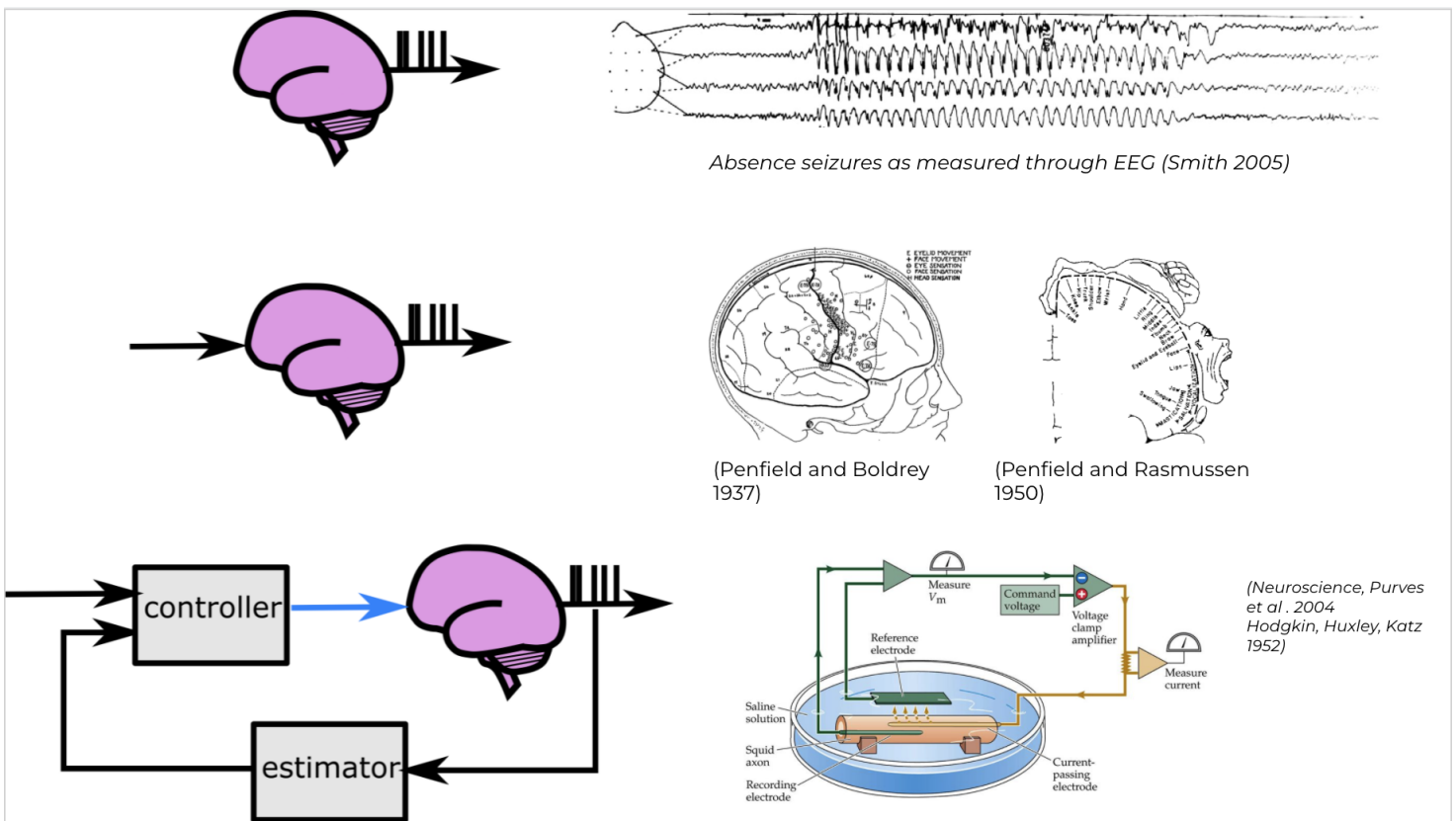# Interventions in neuroscience & causal inference

> 🖊 **50% done:**

▶ ↪Outline
▶ ↪See also

# 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.

(close to final, but could be significantly cut down / merged with other figure) **Figure: 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).
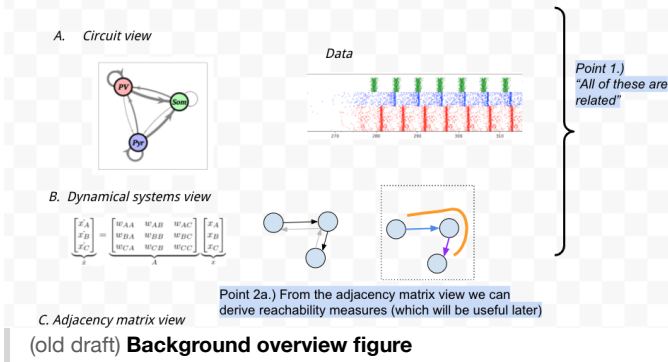
Although theoretical guarantees and algorithms designed to choose among these interventions are often designed for simple models with strong assumptions on properties such as the types of functional relationships that exist in circuits, the visibility and structure of confounding relationships, and noise 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:**

## 2a.) Methods overview
*(goal: introduce language of graphs, adj matrices, dynamical systems, interventions)*

A. *Circuit view*

Data

Point 1.)
"All of these are related"

B. *Dynamical systems view*

$$\begin{bmatrix} x_A \\ x_B \\ x_C \end{bmatrix} = \begin{bmatrix} w_{AA} & w_{AB} & w_{AC} \\ w_{BA} & w_{BB} & w_{BC} \\ w_{CA} & w_{CB} & w_{CC} \end{bmatrix} \begin{bmatrix} x_A \\ x_B \\ x_C \end{bmatrix}$$

Point 2a.) From the adjacency matrix view we can derive reachability measures (which will be useful later)

C. *Adjacency matrix view*

(old draft) **Background overview figure**

▶ ↪ graph for shared v.s. private sources

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]

$$X = XW + E,$$

▶ various implementations

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*

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

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 $\widetilde{W}_{ij} \neq 0$, with the magnitude of $\widetilde{W}_{ij}$ indicating sensitive node $j$ is to a change in node $i$.

This notion of reachability, encoded by the pattern of nonzero entries in $\widetilde{W}$, 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 intro sections ta
```
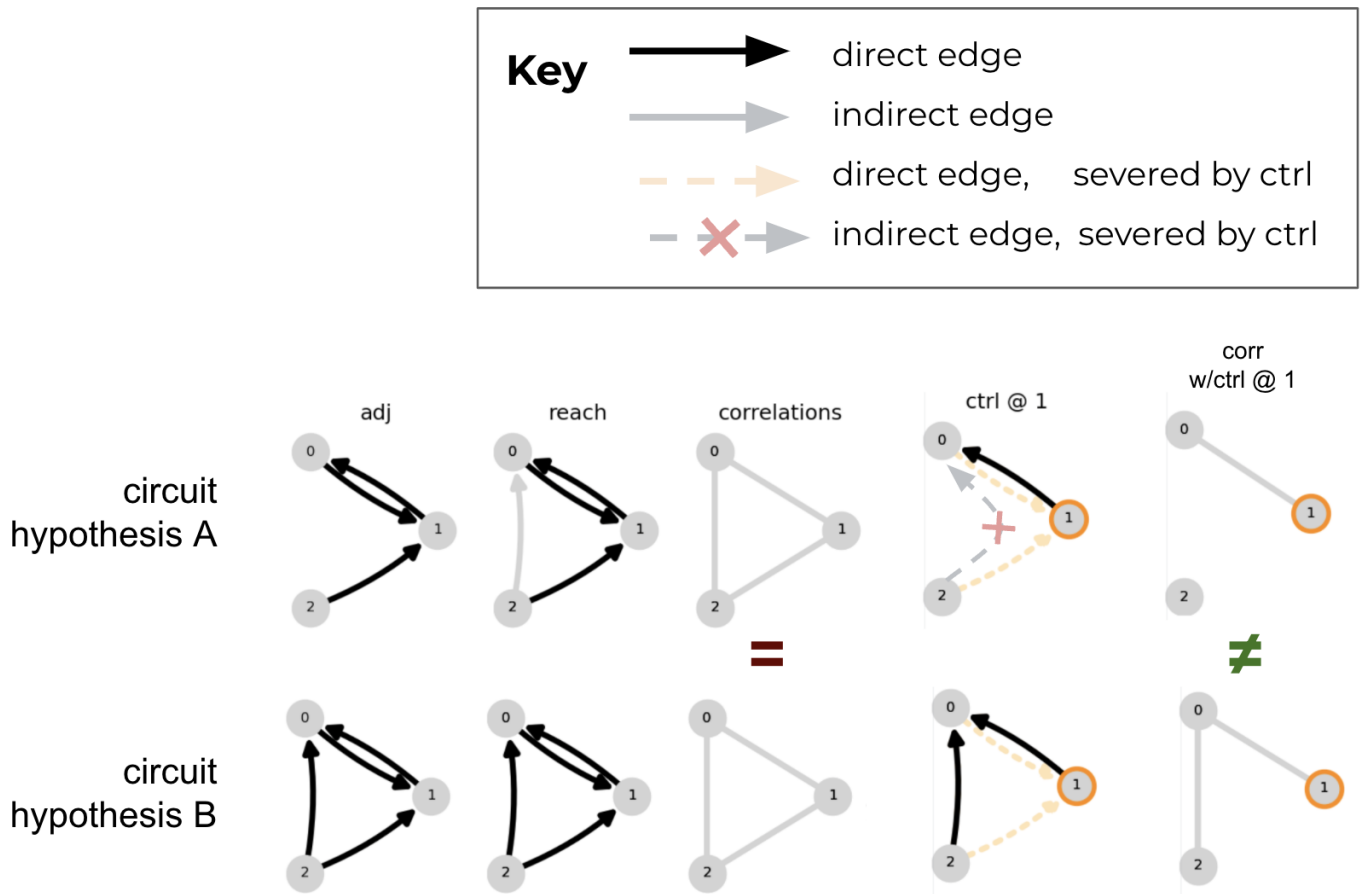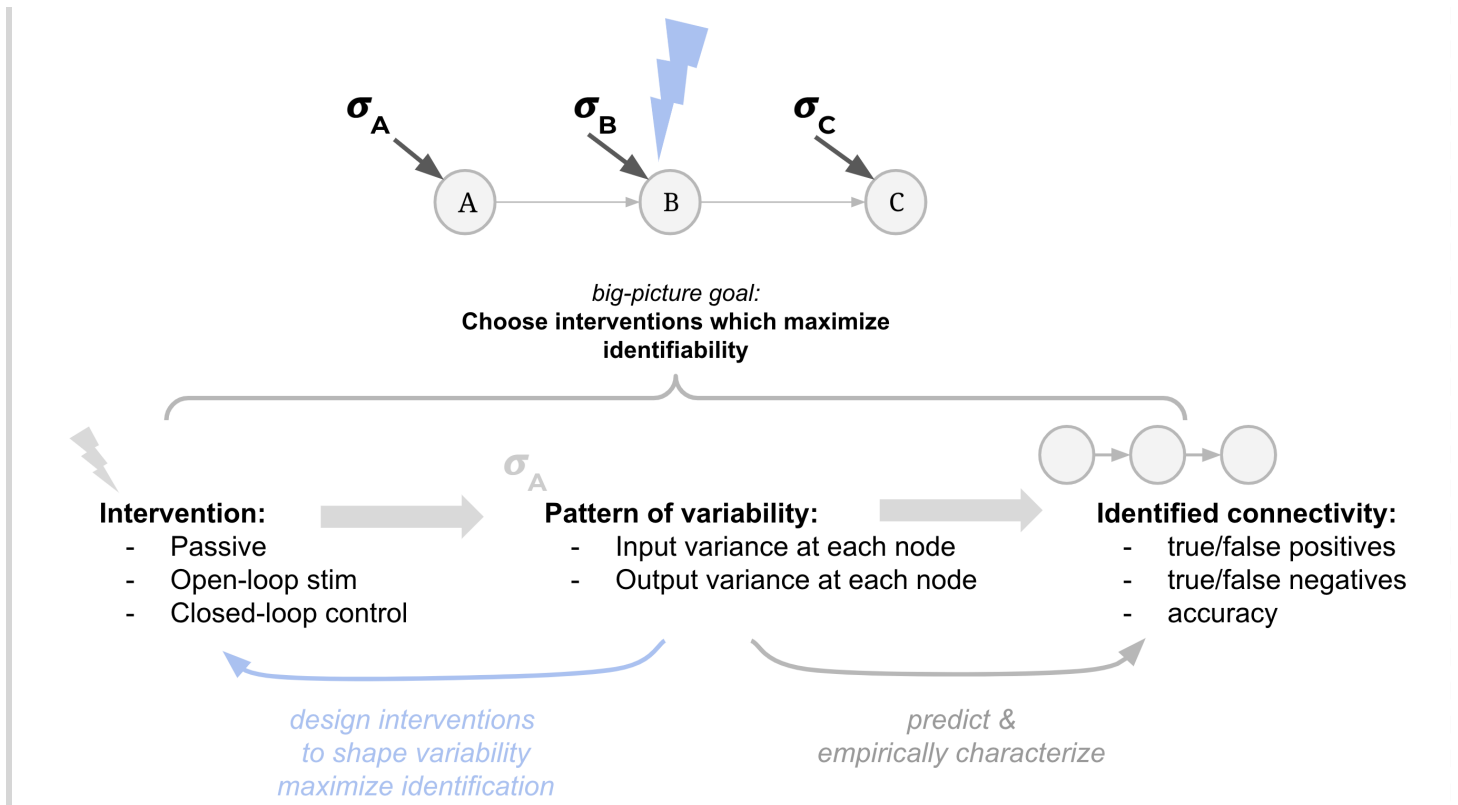
▶ ↪old reachability → ID demo text

(img close to final draft) **Figure DEMO: Applying CLINC to distinguish a pair of circuits**

▶ ↪2,3 circuit versions, straight from code
▶ ↪to do items
▶ ↪see also
▶ ↪more notes
☐ @adam sketch the flow of the argument
☐ @matt to round out writing demo / example walkthrough

# Theory / Prediction

*(Draft overview)*

*big-picture goal:*
**Choose interventions which maximize identifiability**

**Intervention:**
- Passive
- Open-loop stim
- Closed-loop control

**Pattern of variability:**
- Input variance at each node
- Output variance at each node

**Identified connectivity:**
- true/false positives
- true/false negatives
- accuracy

*design interventions to shape variability maximize identification*

*predict & empirically characterize*

## 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

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

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

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[8]

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

where $\widetilde{W} = \sum_{k=0}^{\infty}(W)^k$ denotes the *weighted reachability matrix*, whose $(i,j)^{\text{th}}$ entry indicates the total influence of node $i$ on node $j$ through both direct and indirect connections.[9] 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

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

TODO do a quick matlab simulation to check all of this -- some errors may have been introduced when changing notation

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 $k = 1, \ldots p$, 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.

# Simulation Methods

scope markers:

- ✅ - currently in scope
- 🌑 - want to be in scope, have a head-start
- 🚀 - want to be in scope, would require substantial work
- 🙈 - not intended to be in scope, future work

# Methods overview

- METHODS SUMMARY (high-level): FIRST, what question are we trying to answer
  - then OVERVIEW of methods (pipeline summary)
  - basic components
  - answer we're looking for
  - overall approach
  - key innovative methods
  - assume readers aren't going to pore over the details

▶ ↪overview sections:

# Modeling network structure and dynamics

> ✏️ **60% done**

▶ ↪to do

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

- 1. linear-gaussian v.s. spiking/rate 🌑

The first approach is accomplished with a network of nodes with gaussian noise sources, linear interactions, and linear 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[14].

Neuron dynamics:

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

- 2. contemporaneous vs lagged 🌑
  Additionally we study two domains of interactions between populations; contemporaneous and delay-resolvable connections. These

$$\text{domain} = \begin{cases} \text{contemporaneous}, & \delta_{syn} < \Delta_{sample} \\ \text{delay-resolvable}, & \delta_{syn} \geq \Delta_{sample} \end{cases}$$

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[15] so this temporal precedence clue is lost (although directionality can still be inferred in the presence of intervention).

▶ ↪concept figures
  - in the linear gaussian case we focus on "contemporaneous" domain, for simplicity, then extend to the connections-with-delay case

## Code implementation

Code 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.

*see _network_parameters_table.md for list of relevant parameters*

▶ ↪outline
▶ ↪longer outline
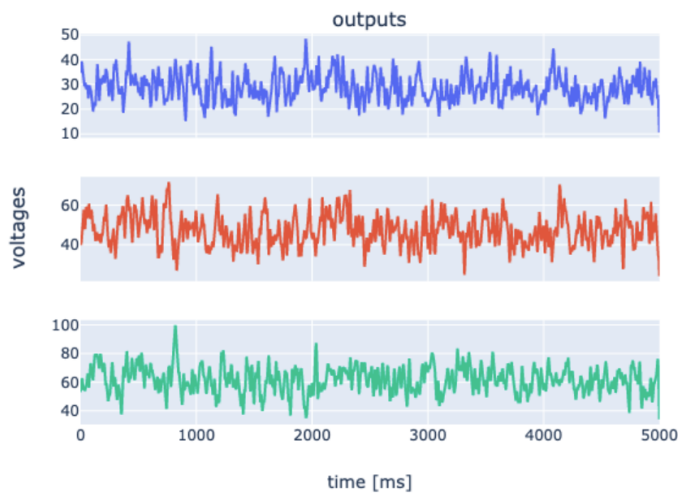▶ see also

## Implementing interventions

✎ **10% done**

▶ ↪ short outline
▶ ↪ long outline

## Extracting circuit estimates

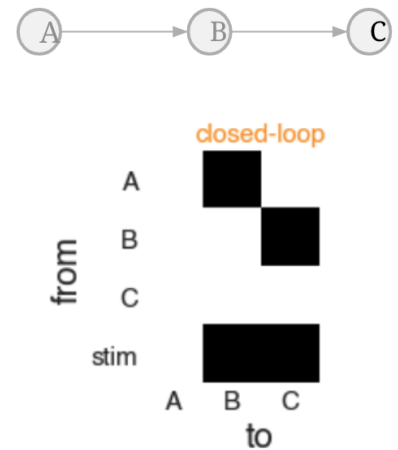✎ **10% done**

*1. Aggregating network data*  *2. Extracting co-dependence*  *3. Thresholding, statistical tests*

▶ ↳outline
▶ ↳longer outline

## Information-theoretic measures of hypothesis ambiguity

✏ **10% done**

*see _steps_of_inference.md for entropy writeup*

# Results

✏ **overall, 40% done**

## Impact of intervention on estimation performance

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

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

▶ ↳notes, see also

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

**Entropy** as a measure of distinguishing circuits

$H(passive) = 0.59$

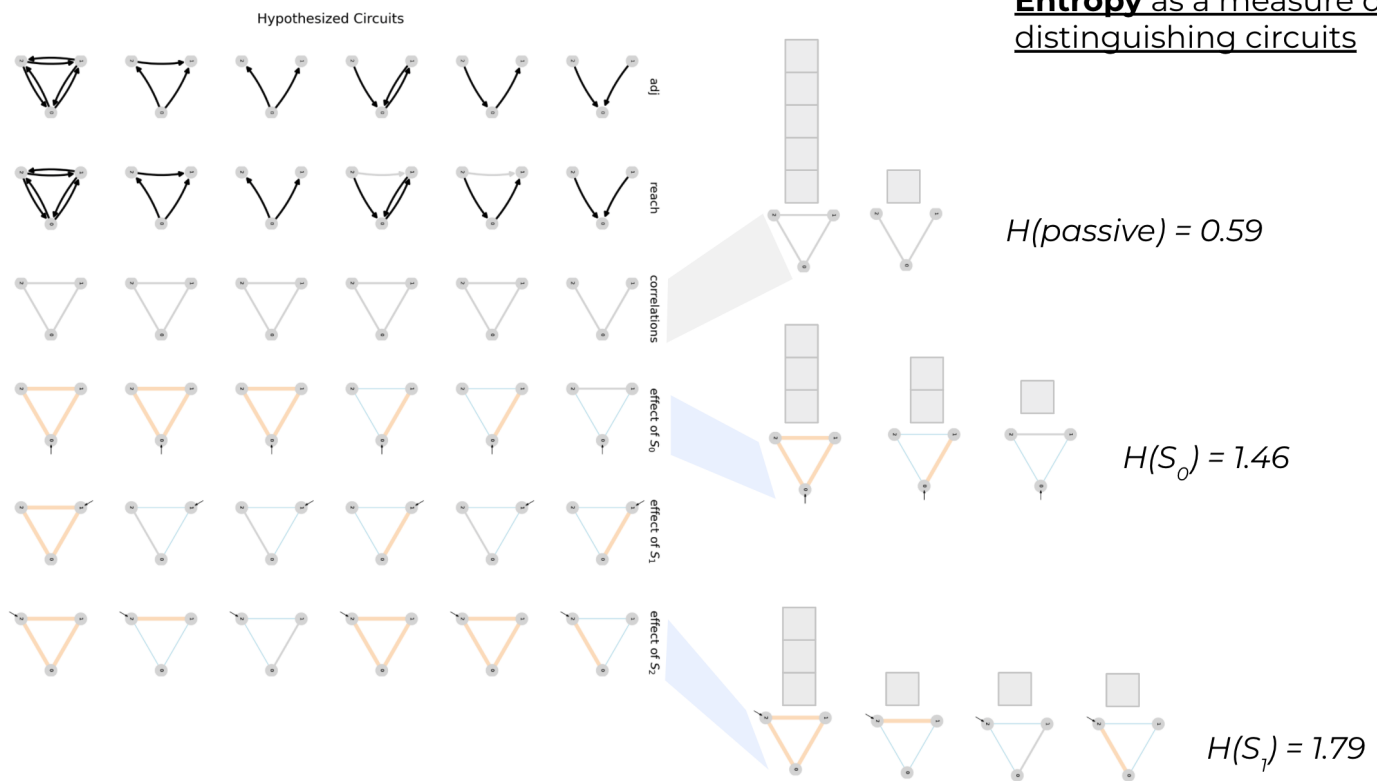$H(S_0) = 1.46$

$H(S_1) = 1.79$

**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[17] edges denote correlations which increase above their baseline value with high variance open-loop input. Thin blue[17:1] 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**[18] 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[19] 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[20]

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

Applying closed-loop to a linear gaussian circuit:

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

- ▶  ↪ Firing rates couple mean and variance
- ▶  ↪ Notes on imperfect control

> ✏️  **reference figvar to emprically 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[22]  `[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 $\text{CoReach}(i, j|S_k)$. 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→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.

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

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

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

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 $i \to j$ exists.

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

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 $\widetilde{W}_{S_k \to i} \approx \widetilde{W}_{S_k \to j}$ (upper right * in fig. variance). If i→j are connected $\widetilde{W}_{S_k \to i} \gg \widetilde{W}_{S_k \to j}$ results in an variance-correlation relationship similar to the "upstream source" case (increasing source variance increases correlation $\frac{dR}{dS_k} > 0$),
while $\widetilde{W}_{S_k \to i} \ll \widetilde{W}_{S_k \to j}$ results in a relationship similar to the "downstream source" case ($\frac{dR}{dS_k} < 0$)[23]

# **Quantitative impact** of parameters
# Well-predicted by ID-SNR
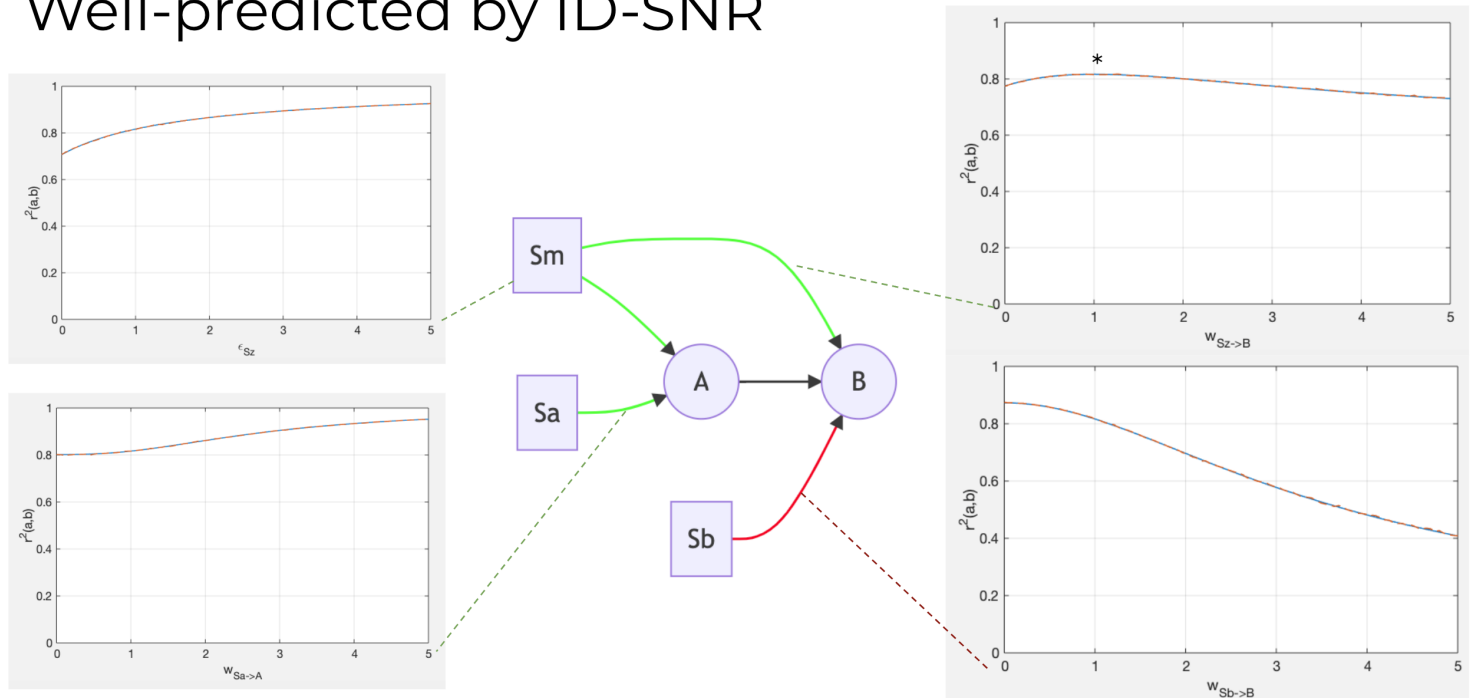
# Closed-loop intervention enables **bidirectional control of correlation**
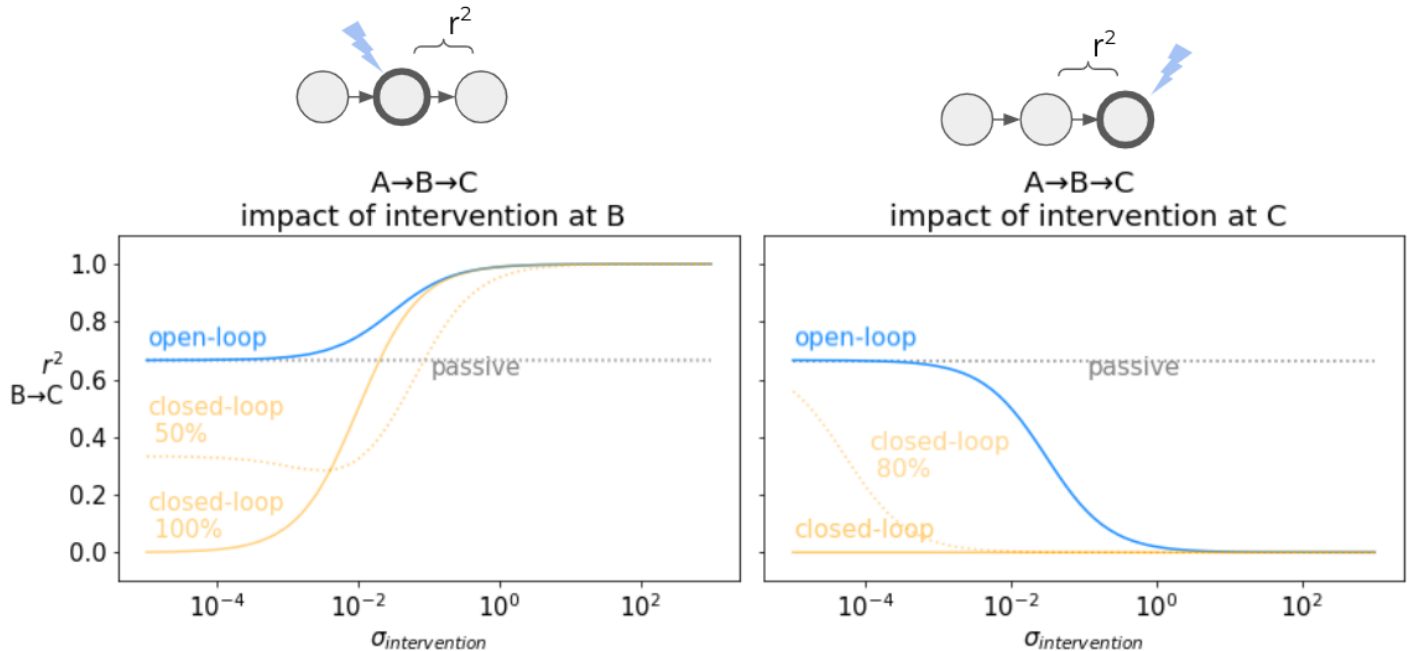*Impact in a linear-gaussian chain, two intervention locations*

**A→B→C**
**impact of intervention at B**

**A→B→C**
**impact of intervention at C**

🚧 (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→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 $\text{Reach}(S_k \to A)/\text{Reach}(S_k \to B)$, 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.

[24]

▶ ↪ additional notes:

🚧

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
    🚧

▶ figure sketches

**Figure DATA: Analysis of simulated circuits suggest stronger intervention facilitates identification with less data** [25]

> ✏️ **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

> ✏️ **needs significant technical work and theory!**

**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)
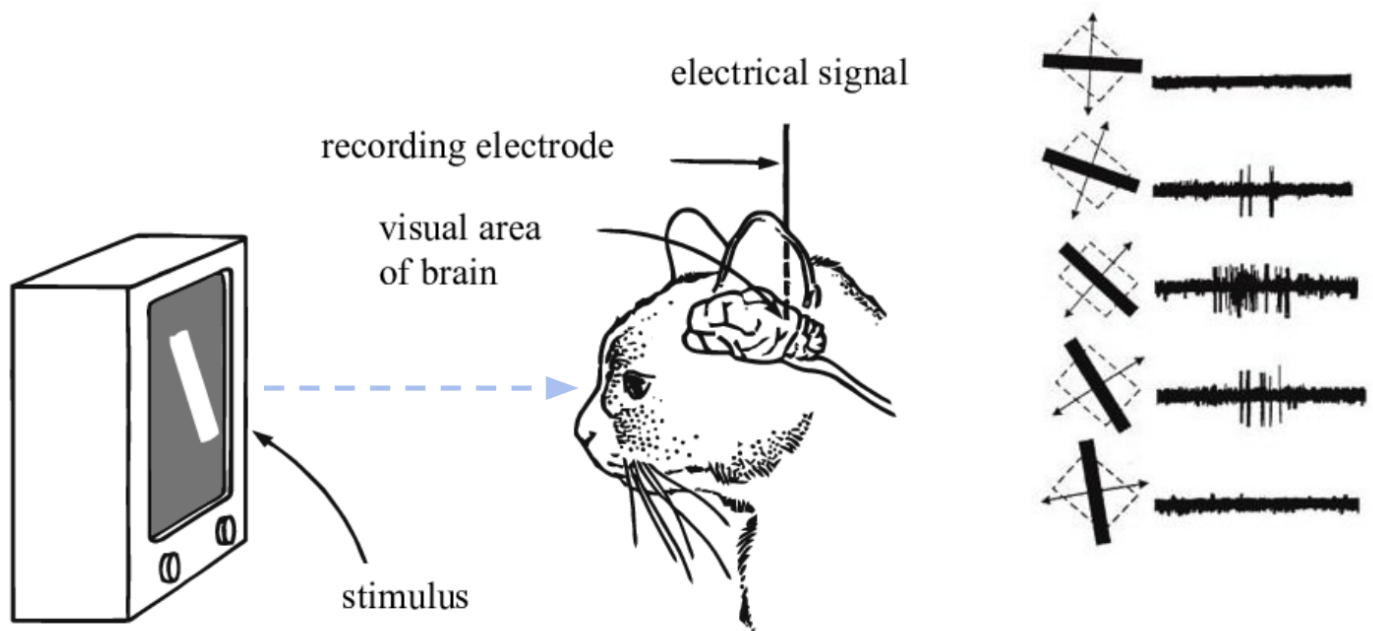
3. probably needs to get more specific sooner, @Adam can fill in ↵

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. 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 $\Sigma = \text{cov}(X) = \mathbb{E}\left[XX^T\right]$ can then be written as $\Sigma = \mathbb{E}\left[(I - W^T)^{-1} EE^T (I - W^T)^{-1}\right] = (I - W^T)^{-1} \text{cov}(E)(I - W^T)^{-T} = (I - W^T)^{-1} \text{diag}(s)(I - W^T)^{-T}$. ↵

9. 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. ↵

10. see causal_vs_expt.md ↵ ↵

11. "The measures implemented are: mutual information, conditional mutual information, Granger causality, and conditional Granger causality (each for univariate and multivariate linear-Gaussian processes). For completeness we have also included Pearson correlation and partial correlation for univariate processes (with a potentially multivariate conditional process)." ↵ ↵

12. a study of problems encountered in Granger causality analysis from a neuroscience perspective ↵ ↵

13. "METHODS FOR STUDYING FUNCTIONAL INTERACTIONS AMONG NEURONAL POPULATIONS" - comes with MATLAB code, discusses time and trial shuffling, decomposing information (synergistic, redundant, independent) ↵ ↵

14. However, depending on overall firing rates and population sizes, this sparse spike-based transmission can be coarse-grained to a firing-rate-based model. ↵

15. the effective \Delta_{sample} would be broadened in the presence of jitter in connection delay, measurement noise, or temporal smoothing applied post-hoc, leading ↵

16. 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. ↵

17. 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 ↵ ↵

18. Figure VAR shows this pretty well, perhaps sink this section until after discussing categorical and quantitative? ↵

19. below the level set by added, independent/"private" sources ↵

20. notably, this is part of the definition of open-loop intervention ↵

21. 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. ↵

22. 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) ↵

23. 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. ↩

24. 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. ↩

25. "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" ↩