

# Closed-Loop Identifiability in Neural Circuits

## Closed-Loop Identifiability in Neural Circuits

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

### 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<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

!!!! - 40% done:

**Estimating causal interactions in the brain.** Many hypotheses about neural circuits are best stated in terms of causal relationships: “changes made in activity in this region of the brain will produce corresponding changes in that downstream region.” Understanding these causal relationships is critical to

---

<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

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 [?].

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 [?] [?] [?], 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 [?], a concept that 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, 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 [?]. 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 [?] [?].

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 [?] [?]. 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 confounding influence [?, Sec. 4.2.3]. 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 [?] or the ability to perturb the system [?] [?]. 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 influence we use the IDTxl toolbox \cite{wollstadt2019idtxl}.]** Despite their mathematical differences, however, previous work has found that cross-correlation-based metrics and information-based metrics tend to produce qualitatively similar results [?] [?, ?].

reviews to read/cite: (todo)

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

- specific to neuro
  - [?]
  - [?]
  - [?]
  - \cite{TODO-knox1981detection}
  - [?]

- [?]
- maybe...
  - [?]
  - [?]

*see also why\_control.md*

## Interventions in neuroscience & causal inference

!!!! - 50% done:

### Outline

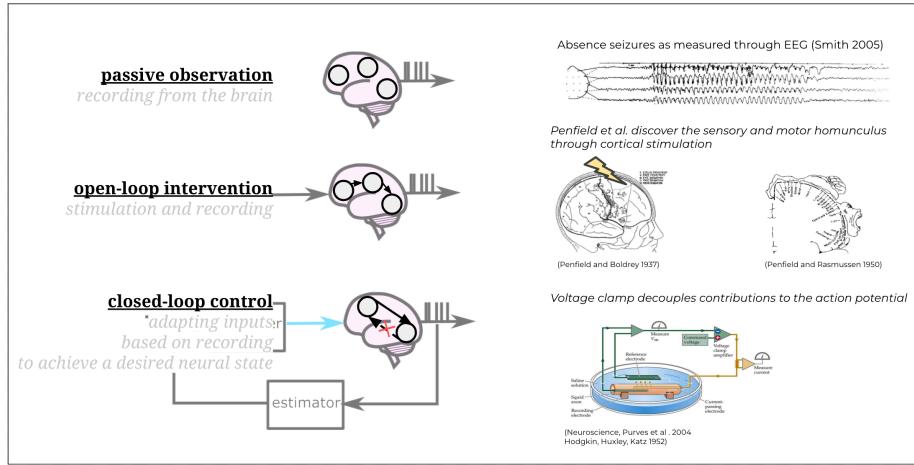
- core idea is that “stronger” interventions lead to “higher inferential power”
  - may mean identifying circuits with less data
  - but may also mean distinguishing circuits which may have been “observationally equivalent” under weaker interventions
- **Highlight that the impact of interventions may 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

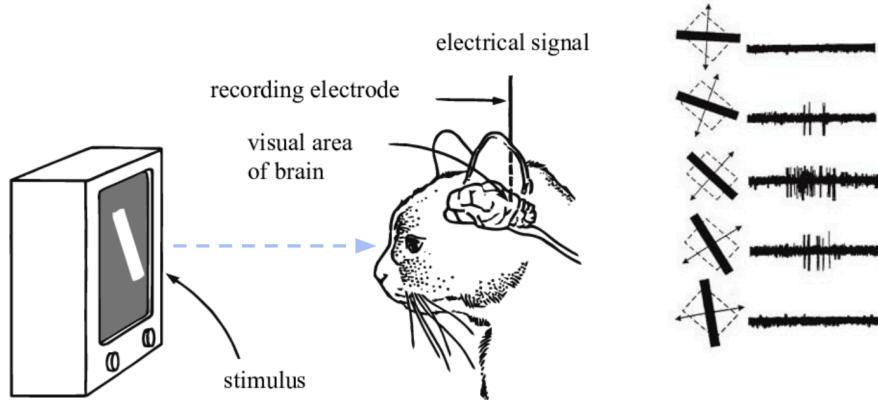
A rich theoretical literature has confirmed the central role of interventions in inferring causal structure from data [?, ?]. Consistent with intuition from neuroscience literature, data in which some variables are experimentally intervened on is typically much more powerful than observational data alone. 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.

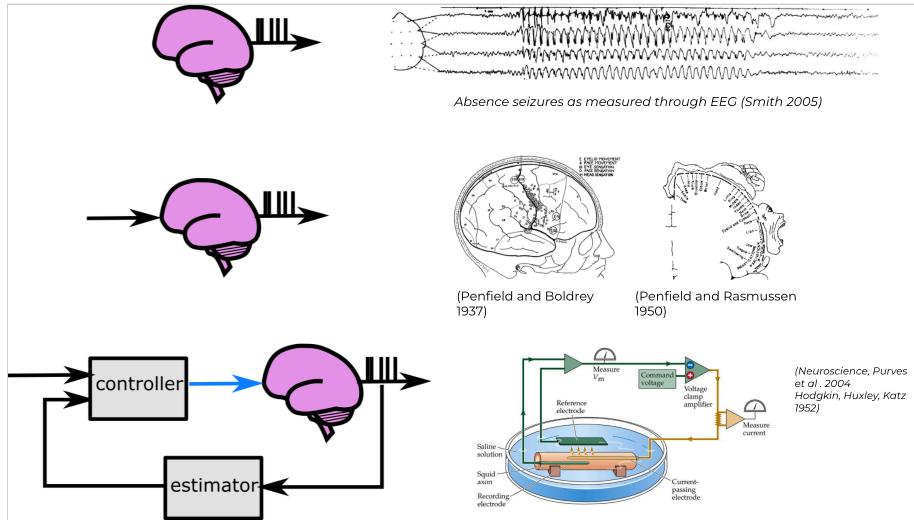


**Figure INTRO: Figure (Interventions in Neuro):** Examples of the role of interventions in discoveries in neuroscience (A) Identifying when a patient is having a seizure, from **passive recordings** alone (B) through systematic **open-loop stimulation experiments**, Penfield was able to uncover the spatial organization of how senses and movement are mapped in the cortex <sup>2</sup> (C) **Feedback control** allows us to specify activity in the brain in terms of outputs. Allows us to reject disturbances, respond to changes (conceptual overview of interacting regions, intervention, DAGs etc.)

prev. figure

<sup>2</sup>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)

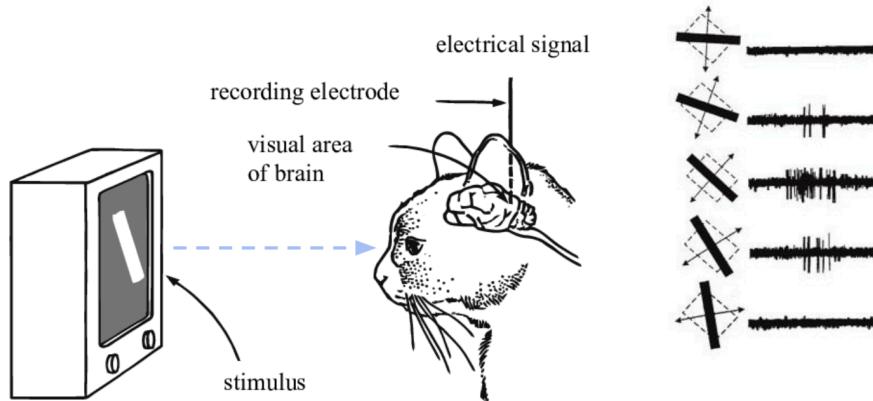




> (close to final, but could be significantly cut down / merged with other figure) **Figure: Examples of the role of interventions in discoveries in neuroscience** (A) Identifying when a patient is having a seizure, from passive recordings alone (B) through systematic open-loop stimulation experiments, Penfield was able to uncover the spatial organization of how senses and movement are mapped in the cortex <sup>3</sup> (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.<sup>4</sup> Interventions on some portions of a system may allow more information about the system's causal structure than interventions in other areas. Interventions

<sup>3</sup>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)



<sup>4</sup>probably needs to get more specific sooner, @Adam can fill in

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<sup>5</sup>. 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 help practitioners design experiments that provide as much scientific insight as possible<sup>6</sup> [?, ?]. Importantly, the necessity and inferential power of interventions is often *algorithm-independent*, in the sense that there exist interventions that reveal causal structure that would be impossible for *any* algorithm to infer from observational data alone [?].

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:

Different mathematical representations of circuits can elucidate different connectivity properties. For example, consider the circuit  $A \rightarrow B \leftarrow C$ . 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<sup>7</sup>.

---

<sup>5</sup>@Adam - make this more precise. talk about spatial, temporal degrees of freedom

<sup>6</sup>@Matt this needs breaking down

<sup>7</sup>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.

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<sup>8</sup>

$$X = XW + E,$$

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  $A \rightarrow B \leftarrow C$ , 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.

---

<sup>8</sup>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. have to resolve / compare these implementations

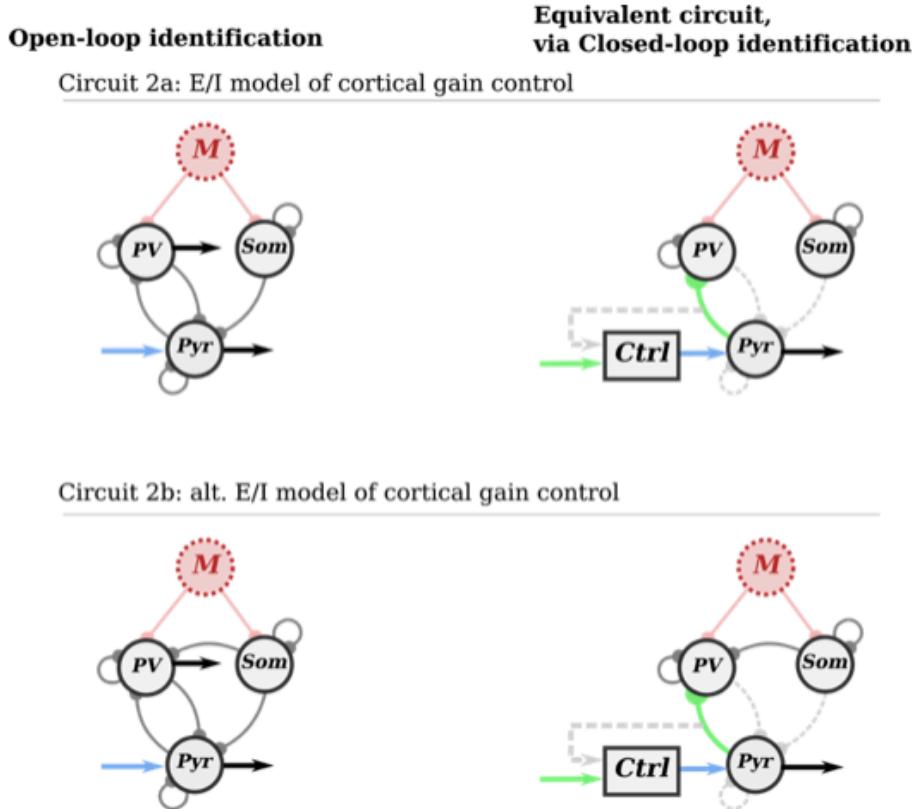
[Matt to Adam --- I like the idea of an example here, but the details will likely need to change once the neighboring intro sections take shape]

!!!! - transition from reachability to 2-circuit ID demo is now in background\_id\_demo.md

old reachability → ID demo text

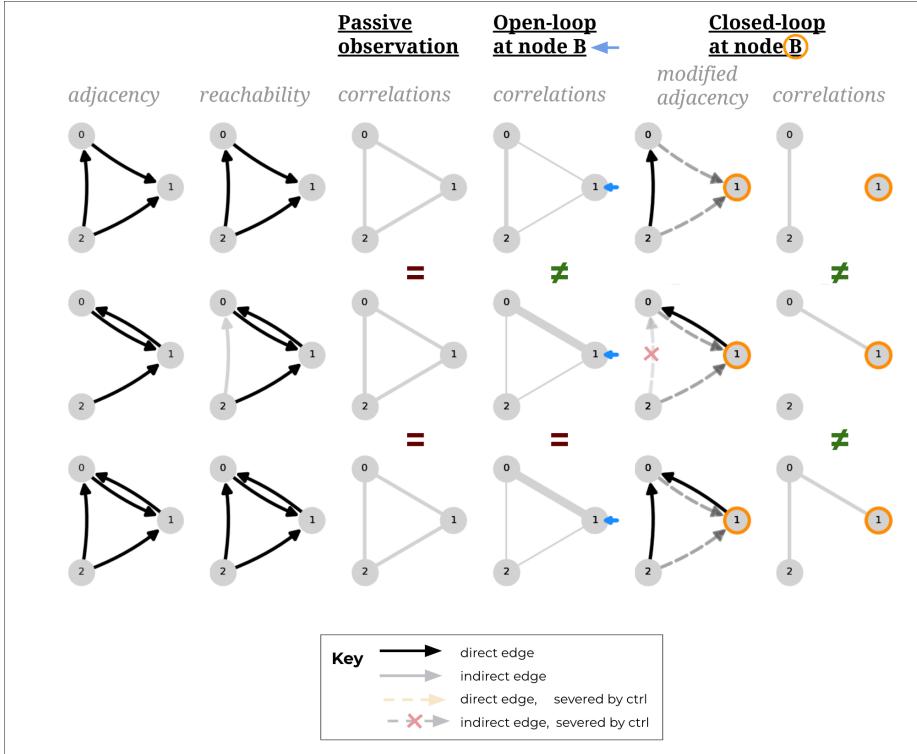
Consider, for example, the hypotheses for cortical gain control in open-loop (Figure BACKGROUND>REPRESENTATION/REACH-1, left column). In both circuit 2a and 2b, PV cells are reachable from the Som cell node ( $\widetilde{W}_{PV \rightarrow Som} \neq 0$ ), since Som activity can influence PV activity 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.

!!!! - 15% done -> much closer now, awaiting reassessment by Adam



> **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 ??: 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>9</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

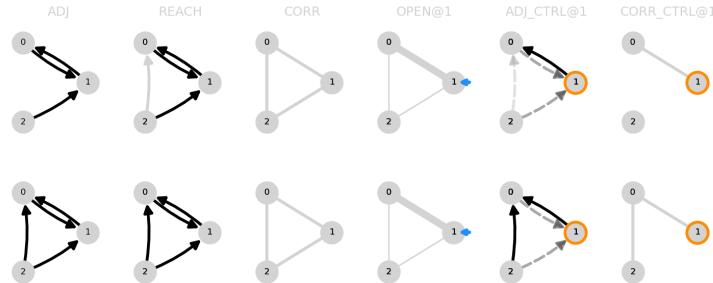
<sup>9</sup>saying “difficult to distinguish” instead of “indistinguishable” here since the magnitudes of the correlations could also be informative with different assumptions

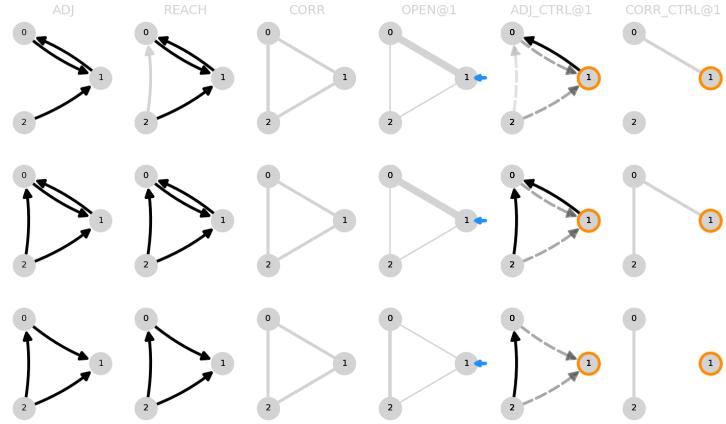
0 in hypotheses B and C, meanwhile, causes the open-loop stimulation at node 1 to *increase* the observed correlation between nodes 1 and 0. An experimenter can thus distinguish between hypothesis A and the other two hypotheses by applying open-loop control and observing the resulting pattern of correlations (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 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

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

2,3 circuit versions, straight from code





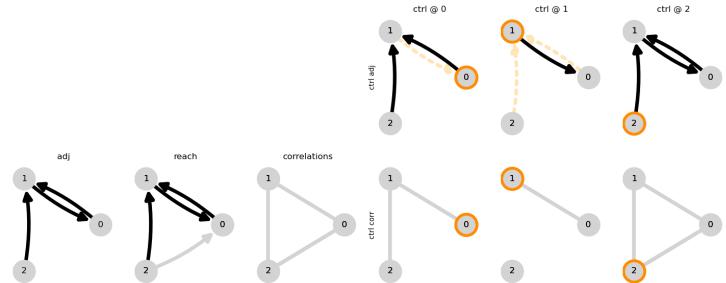
> 3 circuit walkthrough, walkthrough will all intervention locations might be appropriate for the supplement

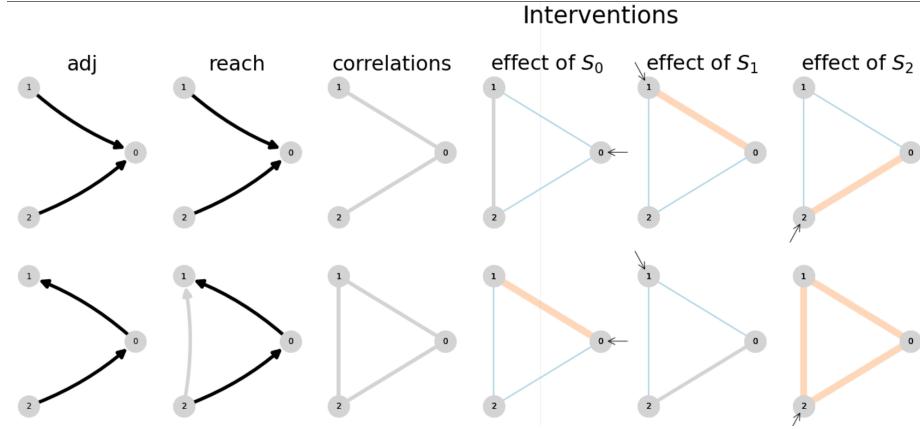
to do items

- 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/generated/networkx.algorithms.](https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms)

see also

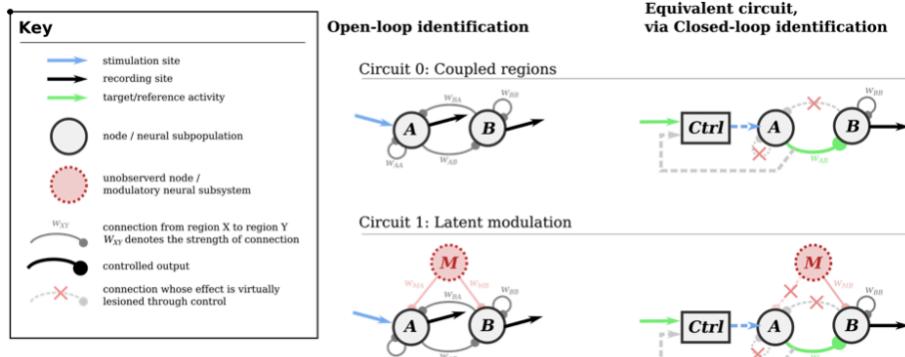
more inspiration: - Combining multiple functional connectivity methods to improve causal inferences - Advancing functional connectivity research from association to causation - Fig1. of “Systematic errors in connectivity”





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

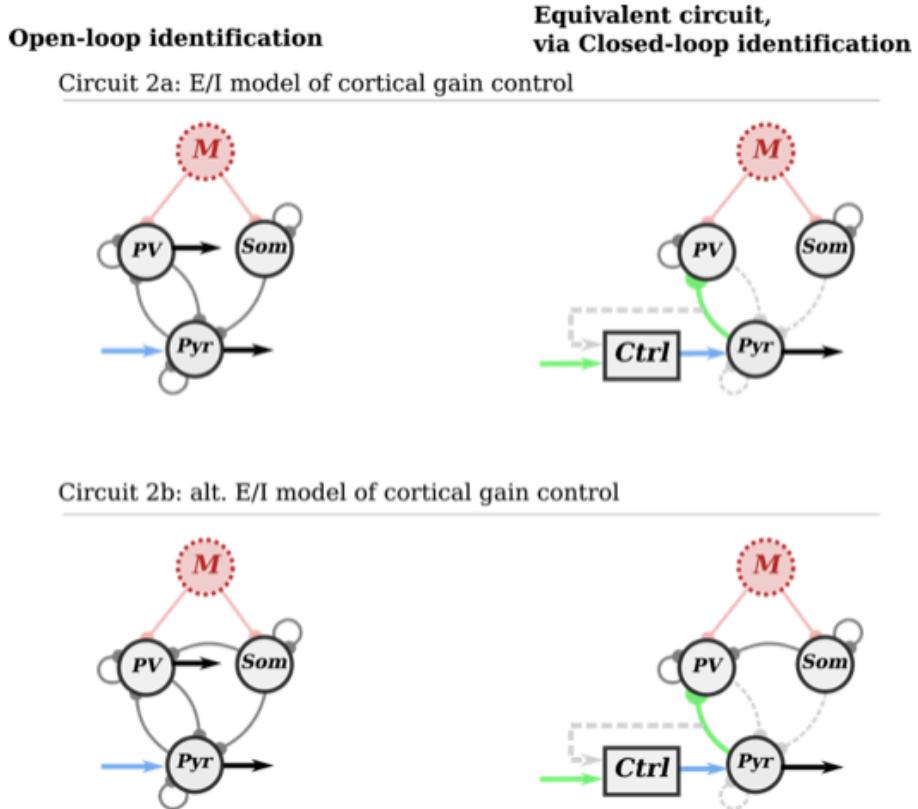


**Figure 11: Closed-loop control compensates for inputs to a node in simple circuits:** The left column shows a simple circuit and recording and stimulation sites for an open-loop experiment. The right column shows the functional circuit which results from closed-loop control of the output of region 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 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.

more notes

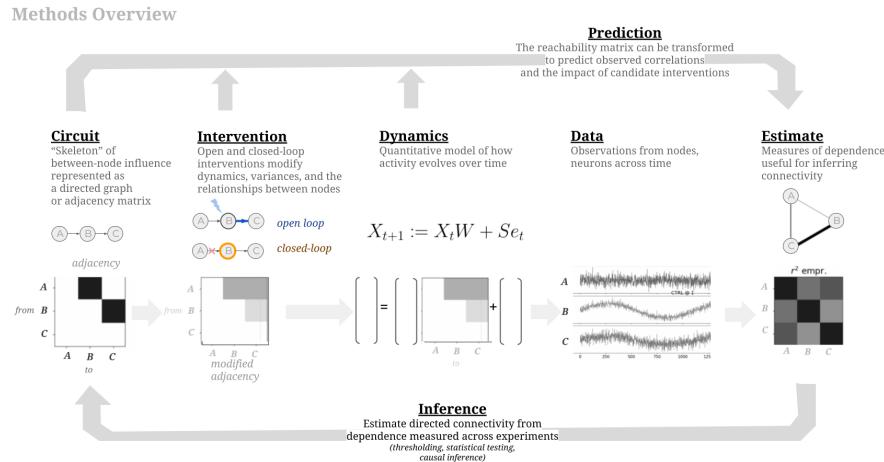
probably want - two circuits which look clearly different - ! but which have equivalent reachability - possibly with reciprocal connections - posssibly with common modulation

- 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
  - $\text{IDSNR}_{ij}$  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  $\text{IDSNR}$  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>10</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  $\widetilde{W} = \sum_{k=0}^{\infty} (W^k)$  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.<sup>11</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

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

==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  $k = 1, \dots, p$ , 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.

---

<sup>10</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  $\Sigma = \mathbb{E}[(I - W^T)^{-1} E E^T (I - W^T)^{-T}] = (I - W^T)^{-1} \text{cov}(E)(I - W^T)^{-T} = (I - W^T)^{-1} \text{diag}(s)(I - W^T)^{-T}$ .

<sup>11</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.

## Simulation Methods

```
@ import "methods0_0_overview.md"  
@ import "methods0_1_simulations.md"
```

### Implementing interventions

!!!! - 10% done

short outline

be sure to reference / not reinvent background\_intervention\_causa  
l\_inf.md

#### passive

- baseline drive comes from independent, “private”, noise sources

#### open-loop

- variance modulated
- replay

#### closed-loop

- perfect
- emulated partial
- practical PID
- :rocket: stimulus-conditional transfer entropy

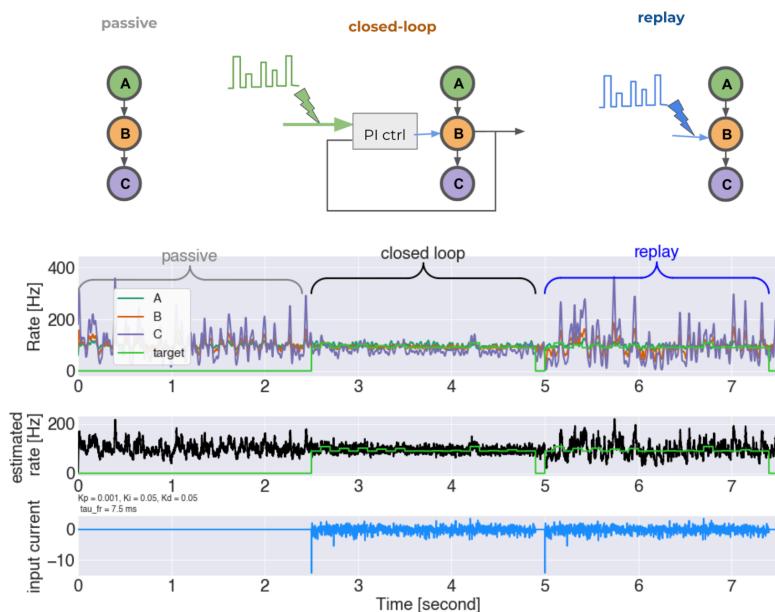
long outline

- passive observation
- open-loop stimulation
  - simulated as direct current injection
  - but uniform across a population
  - ( see Kyle Johnsen’s cleosim toolbox for more detailed simulation of stimulation )
  - closed-loop replay ?
- closed-loop stimulation
  - approaches for control
    - \* going with “model-free” PID control of output rates
  - comparison to randomization in traditional experiment design<sup>12</sup>
  - controller strength
    - \* gain

---

<sup>12</sup>see causal\_vs\_expt.md

- \* bandwidth
- controller delay
- additional stimulation factors (open- & closed-loop)
  - click to expand
  - **stimulus location**
    - \* single-site
    - \* multi-site
    - \* location relative to features of network
      - in-degree/out-degree
      - upstream/downstream of hypothesized connection
  - stimulus intensity
    - \* expected mean output rate
    - \* frequency content

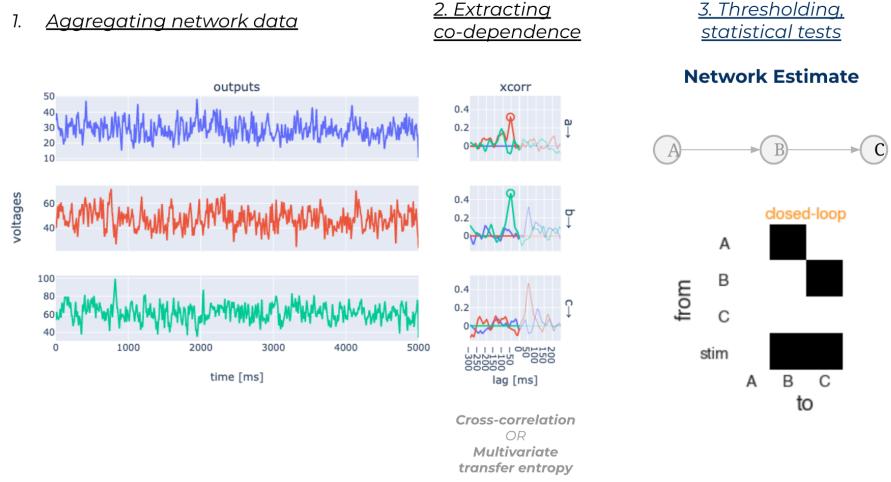


## Extracting circuit estimates

!!!! - 10% done

*refer to methods overview figure*

outline

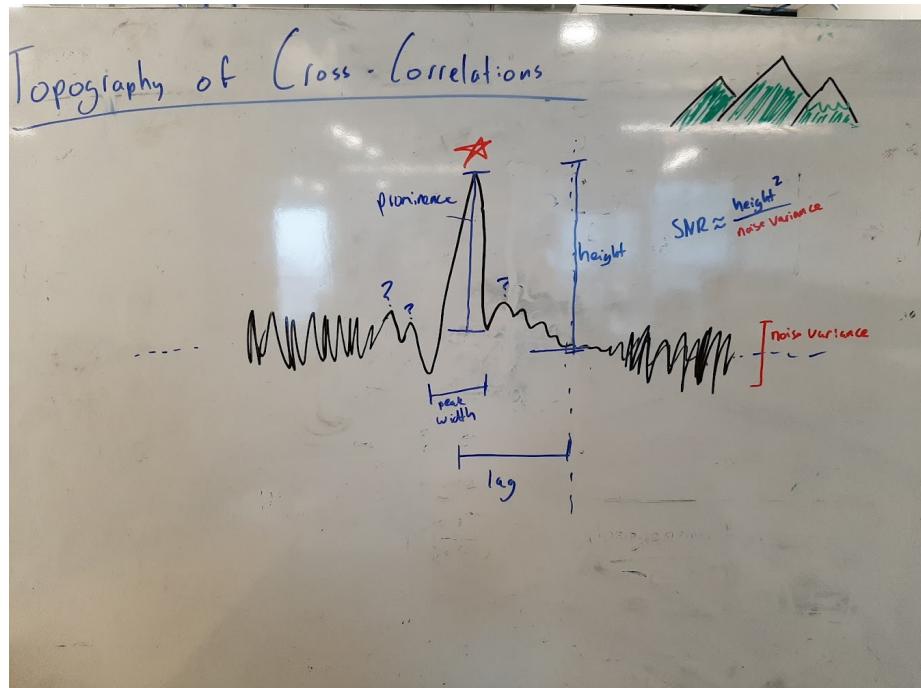


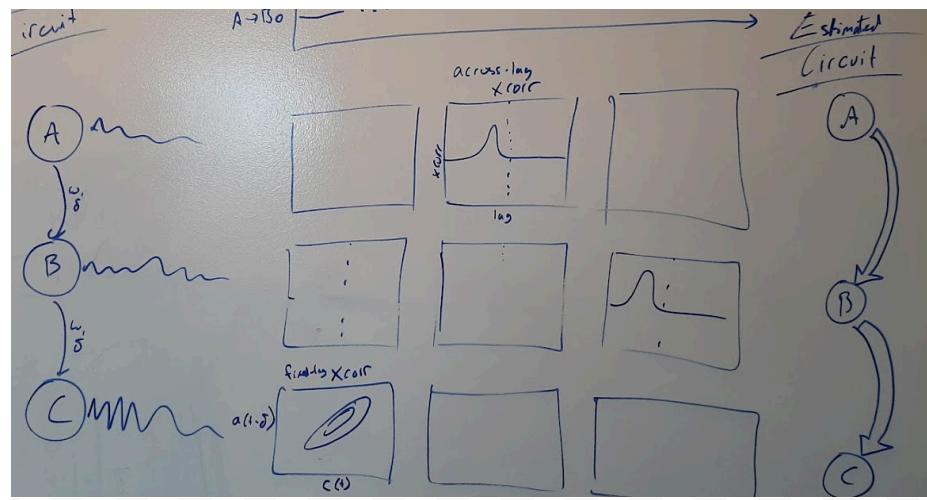
be sure to reference / not reinvent section: **Inferring causal interactions from time series.** in background\_causal\_network\_id.md

## Outputs of network

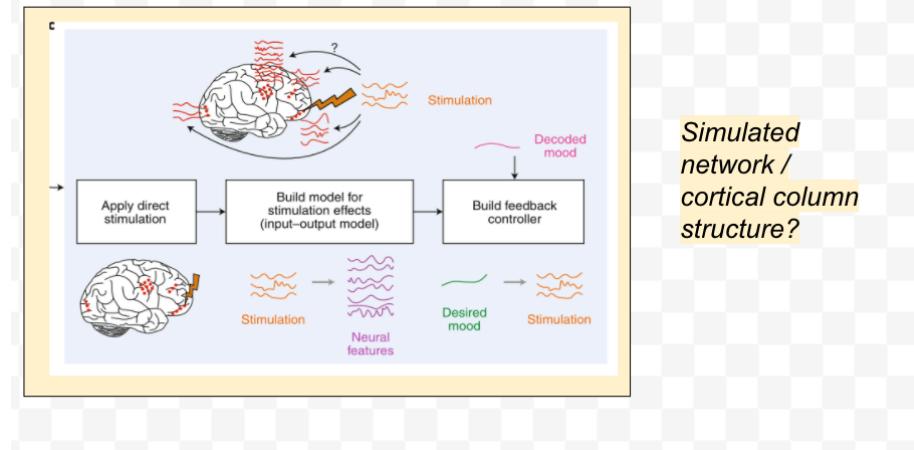
### What is cross-correlation

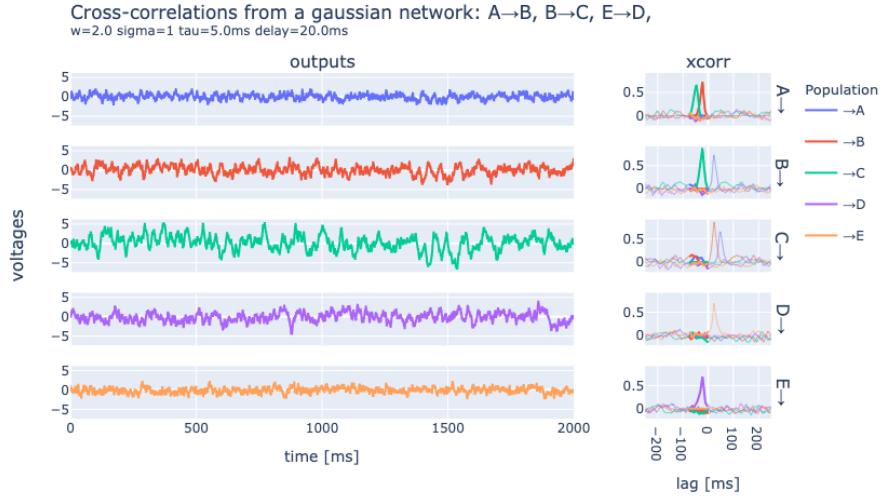
see also





**4a.) Pipeline for inferring circuits with intervention**





**Figure PIPELINE:** Process of detecting connections in a network model

longer outline

- map of techniques available for inference
- see “Assessing the Significance of Directed and Multivariate Measures of Linear Dependence Between Time Series”, code<sup>13</sup> and Unifying Pairwise Interactions in Complex Dynamics
- bivariate v.s. multivariate
- conditioning

```
> - same signals past
> - other signals
> - on stimulus
    • measures of dependence
> - correlation
    > - partial correlation (conditioning)
    > - time-lagged cross-correlation
> - granger causality
> - mutual information
> - transfer entropy
```

---

<sup>13</sup>“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).”

### **lagged cross-correlation**

- connection to / equivalence with Granger Causality (GC)
- review of GC in neuro
- requisite assumptions
- limitations of GC <sup>14</sup>
- xcorr features
- peak-SNR
- prominence
- time of peak
- window of time-lags considered for direct connections
- some multiple of expected synaptic delay

### **multivariate transfer entropy (muTE)**

- advantages above usual GC approach

### **statistical testing**

- *for muTE, handled by IDTxl*
- includes appropriate multiple-comparison testing

### **Quantifying successful identification**

- binary “classification” metrics
- accuracy, F1 score (Wang & Shafechi 2019)
- AUC (Pastore)
- Jaccard index (Lepage, Ching, and Kramer 2013)
- true/false positives, true/false negatives
- graded metrics (*not a core focus here*)
- distance between identified connection strength and ground-truth

> - MSE [(Lepperod et al. 2018)](<https://www.biorxiv.org/content/10.1101/463760v2> )

- error in output reconstruction
- *relevant “negative control” for comparison (?)*
- identified connectivity for random network?
- some shuffled data-surrogate procedure? <sup>15</sup>
- *relevant “positive control” for comparison (?)*

## **Information-theoretic measures of hypothesis ambiguity**

!!!! - 10% done

---

<sup>14</sup>a study of problems encountered in Granger causality analysis from a neuroscience perspective

<sup>15</sup>“METHODS FOR STUDYING FUNCTIONAL INTERACTIONS AMONG NEURONAL POPULATIONS” - comes with MATLAB code, discusses time and trial shuffling, decomposing information (synergistic, redundant, independent)

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

## Results

!!!! - overall, 40% done

### **Impact of intervention on estimation performance**

@ import “/section\_content/results1\_impact\_of\_intervention.md”

### **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