Closed-Loop Identifiability in Neural Circuits

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 -> closer now, awaiting some neuro-writing and status reassessment by Adam

Many hypotheses about neural circuits are phrased in terms of causal relationships: "will changes in activity to this region of the brain produce corresponding changes in another region?" Understanding these causal relationships is critical to both scientific understanding and to developing effective therapeutic interventions, which require knowledge of how potential therapies will impact brain activity and patient outcomes.

A range of mathematical and practical challenges make it difficult to determine

[1] 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

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

Proposal text to draw from:

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

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

Most applications of closed-loop control to neuroscience to date have used "activity-guided / responsive / triggered stimulation" wherein a predesigned stimulus is delivered in response to a detected event. For example, in (Krook-Magnuson et al. 2013) the authors detect seizure activity from spiking and local field potential features to trigger a pulse-train of inhibitory optogenetic stimulation which interrupts the seizure. While this is an effective approach for many applications, these types of closed-loop experiments should be distinguished from closed-loop with ongoing feedback such as dynamic clamp. In these feedback control approaches parameters of stimulation are adjusted on much faster timescales in response to measured activity. For dynamic clamp experiments, this low-latency ongoing feedback control allows experimenters to deliver currents which mimic virtual ion channels which would be implausible with triggered predesigned stimulation. These approaches provide additional

precision in being able to drive activity patterns, but also come with increased algorithmic and hardware demands. For the rest of this document, we will use "closed-loop control" or "feedback control" to refer to this second, more specific class of approaches.

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

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

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

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

Our work initially focuses on measures of directional interaction that are based on lagged correlations [**?**]. These metrics look at the correlation of time series collected from pairs of nodes at various lags and detect peaks at negative time lags. Such peaks could indicate the presence of a direct causal relationship – but they could also stem from indirect causal links or hidden confounders [**?**]. In these bivariate correlation methods, it is thus necessary to consider

patterns of correlation between many pairs of nodes in order to differentiate between direct, indirect, and confounding relationships [**?**]. This distinguishes these strategies from some multivariate methods that "control" for the effects of potential confounders. While cross-correlation-based measures are generally limited to detecting linear functional relationships between nodes, their computational feasibility makes them a frequent metric of choice in experimental neuroscience work [**?**] [**?**] [**?**].

Other techniques detect directional interaction stemming from more general or complex relationships. Information-theoretic methods, which use information-based measures to assess the reduction in entropy knowledge of one variable provides about another, are closely related to Granger causality [**?**] [**?**]. The *transfer entropy* $T_{X \to Y}(t) = I(Y_t : X_{<t} \mid Y_{<t})$ 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 $Y_{<t}$) [**?**]. 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 potential confounding influence [**?**, Sec. 4.2.3]. Conditional transfer entropy can thus be interpreted as the amount of information present in Y that is not contained in the past of X, the past of Y, or the past of other variables Z.

To quantify the strength of causal interactions, information-theoretic and transfer-entropy-based methods typically require knowledge of the ground truth causal relationships that exist [**?**] or an ability to perturb the system [**?**] [**?**]. In practice, these quantities are typically interpreted as "information transfer," and a variety of estimation strategies and methods to automatically select the conditioning set (i.e., the variables and time lags that should be conditioned on) are used (e.g., [**?**]). Multivariate conditional transfer entropy approaches using various variable selection schemes can differentiate between direct interactions, indirect interactions, and common causes, but their results depend on choices such as the binning strategies used to discretize continuous signals, the specific statistical tests used, and the estimator used to compute transfer entropy [**?**]. `[If we end up making the jump to IDTxl in our results: In our empirical results using transfer-entropy-based notions of directional influence we use the IDTxl toolbox \cite{wollstadt2019idtxl}.]` However, despite their mathematical differences, previous work has found that cross-correlation-based metrics and information-based metrics tend to produce qualitatively similar results, with similar patterns of true and false positives [**?**].

Interventions in neuroscience & causal inference

!!!! - 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.

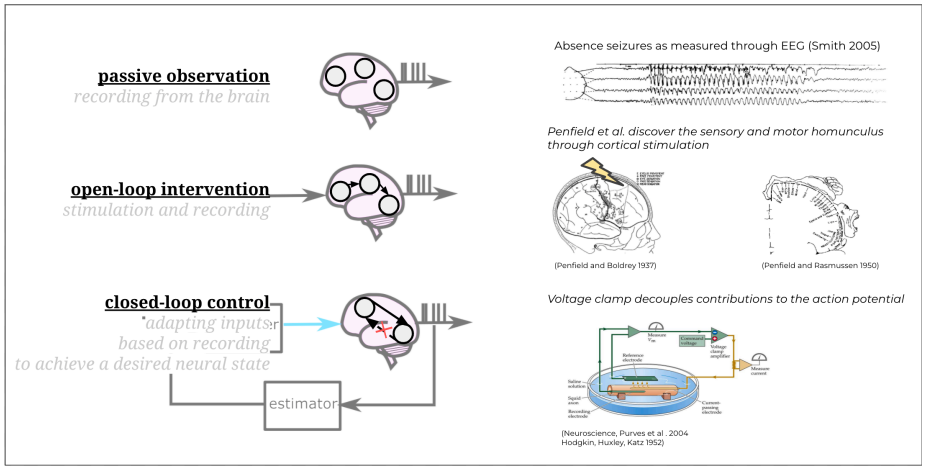
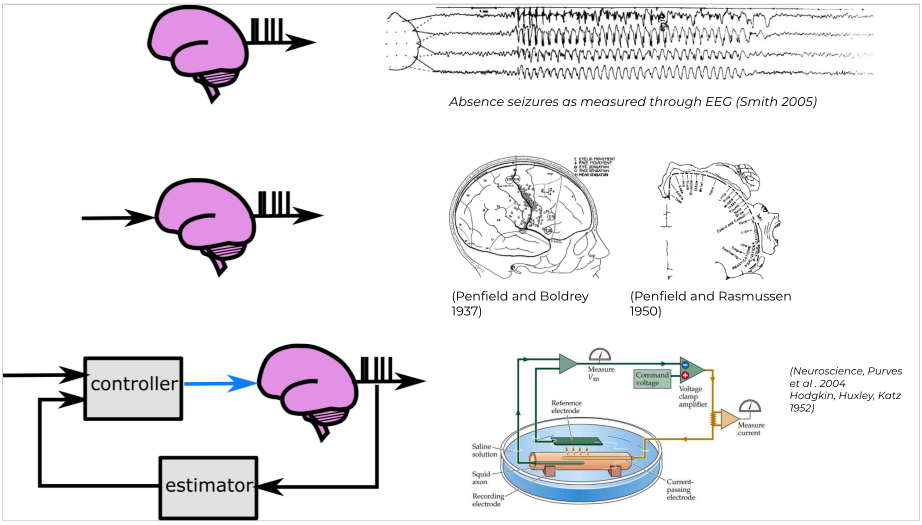


Figure INTRO: Figure (Interventions in Neuro): Examples of the role of interventions in discoveries in neuroscience (A) Identifying when a patient is having a seizure, from **passive recordings** alone (B) through systematic **open-loop stimulation experiments**, Penfield was able to uncover the spatial organization of how senses and movement are mapped in the cortex [2] (C) **Feed-**

[2] Another great example of open-loop mapping: Hubel, D.H., Wiesel, T.N.: Receptive fields

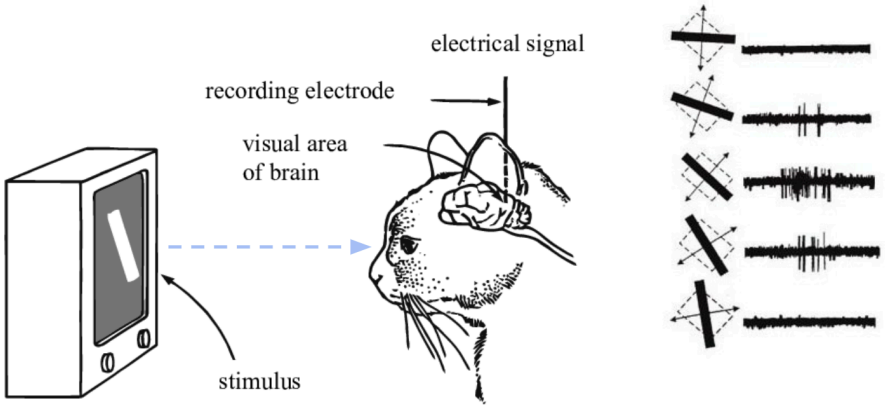
back control allows us to specify activity in the brain in terms of outputs. Allows us to reject disturbances, respond to changes (conceptual overview of interacting regions, intervention, DAGs etc.)

prev. figure



> (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 [3] (C) **Feedback control** allows us to specify activity in the brain in terms of outputs. Allows us to

of single neurones in the cat's striate cortex. The Journal of physiology 148(3), 574–591 (1959)



[3]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)

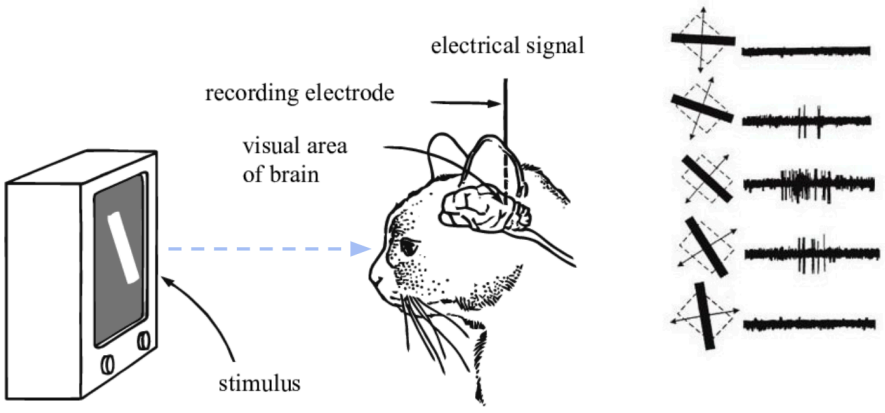
reject disturbances, respond to changes

Certain interventions can provide much more inferential power than others.[4] 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 [5]. 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[6] [**?**, **?**]. 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



[4]probably needs to get more specific sooner, @Adam can fill in
[5]@Adam - make this more precise. talk about spatial, temporal degrees of freedom
[6]@Matt this needs breaking down

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 $A \to B \leftarrow C$. 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[7].

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[8]

$$X = XW + E,$$

!!!! - 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

[7] 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.

[8] 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

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^{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 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 ($\widetilde{W}_{PV \to Som} \neq 0$), 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.

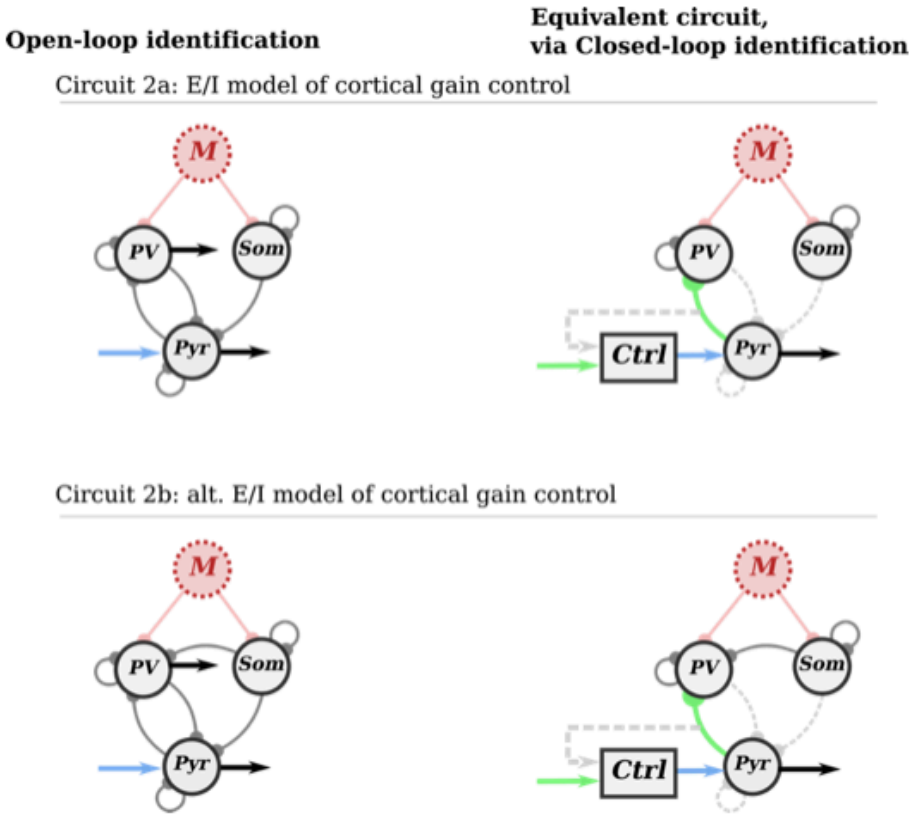
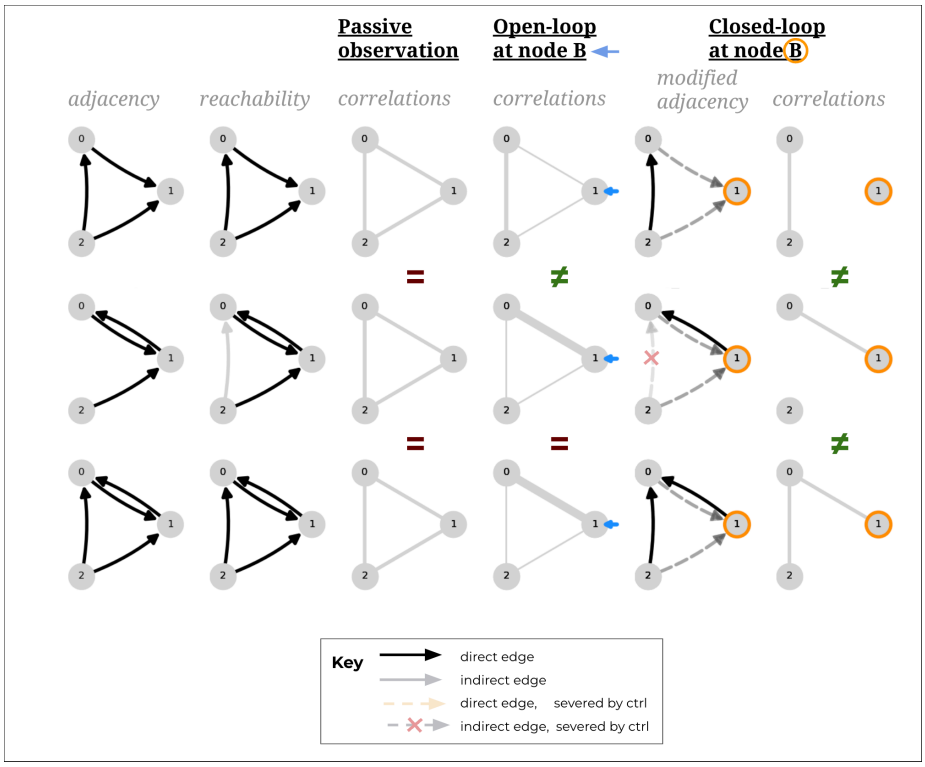


Figure BACKGROUND>REPRESENTATION/REACH-1: Closed-loop control allows for two circuit hypotheses to be distinguished. Two hypothesized circuits for the relationships between pyramidal (Pyr, excitatory), parvalbumin-positive (PV, inhibitory), and somatostain-expressing (Som, inhibitory) cells are shown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

!!!! - 15% done -> much closer now, awaiting reassesment 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 $\widetilde{W}$ in circuit B an *indirect* connection exists through the path $2 \to 1 \to 0$ (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[9] 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

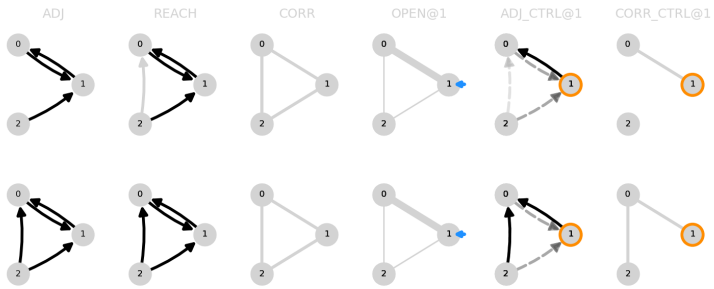
[9]saying "difficult to distinguish" instead of "indistinguishable" here since the magnitudes of the correlations could also be informative with different assumptions

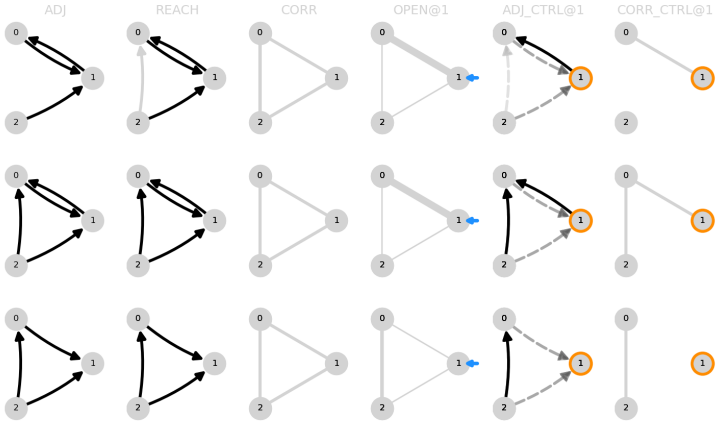
0 in hypotheses B and C, meanwhile, causes the open-loop stimulation at node 1 to *increase* the observed correlation between nodes 1 and 0. An experimenter can thus distinguish between hypothesis A and the other two hypotheses by appling open-loop control and observing the resulting pattern of correlations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





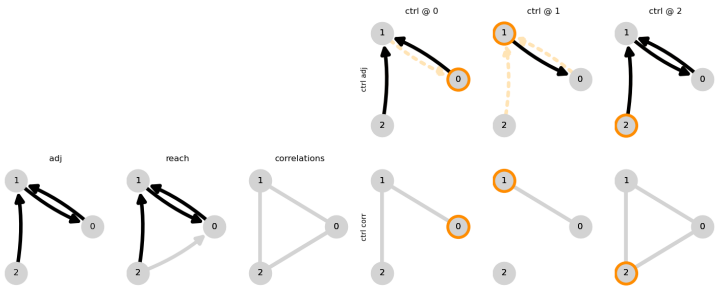
> 3 circuit walkthrough, walkthrough will all intervention locations might be appropriate for the supplement

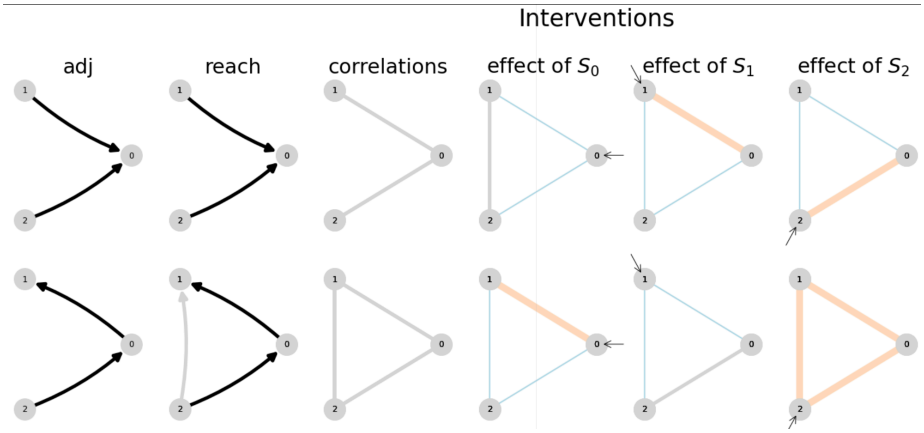
to do items

- ☐ 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.s

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 mutliple hypotheses

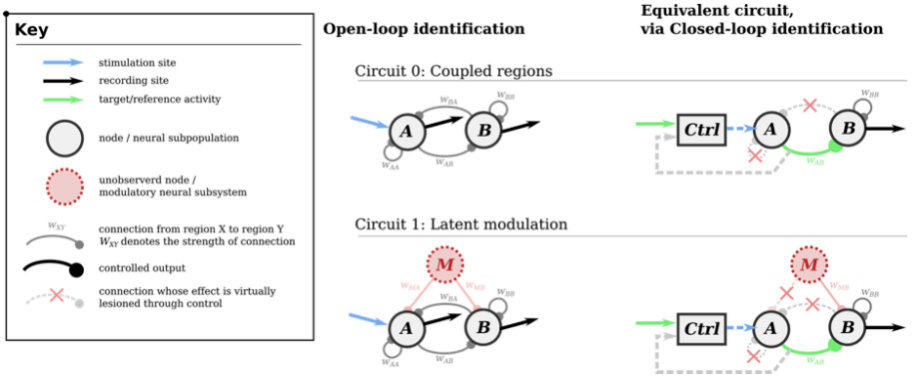


Figure 11: Closed-loop control compensates for inputs to a node in simple circuits: *The left column shows a simple circuit and recording and stimulation sites for an open-loop experiment. The right column shows the functional circuit which results from closed-loop control of the output of region A.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

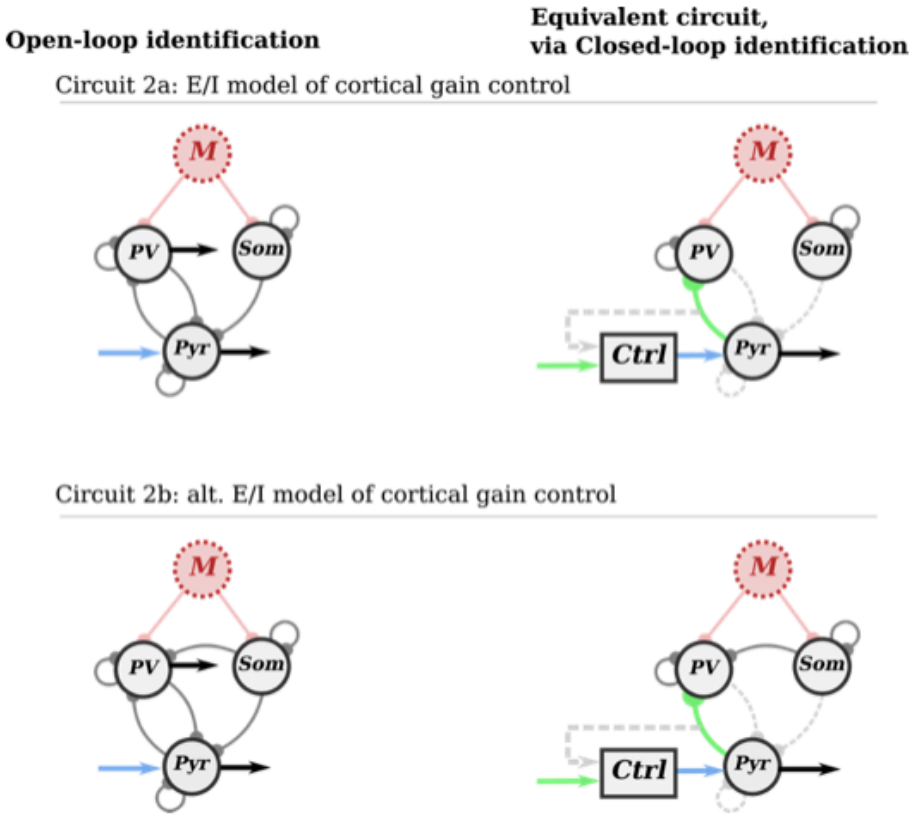


Figure 12: Closed-loop control allows for two circuit hypotheses to be distinguished. Two hypothesized circuits for the relationships between pyramidal (Pyr, excitatory), parvalbumin-positive (PV, inhibitory), and somatostain-expressing (Som, inhibitory) cells are shown in the two rows. Dashed lines 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

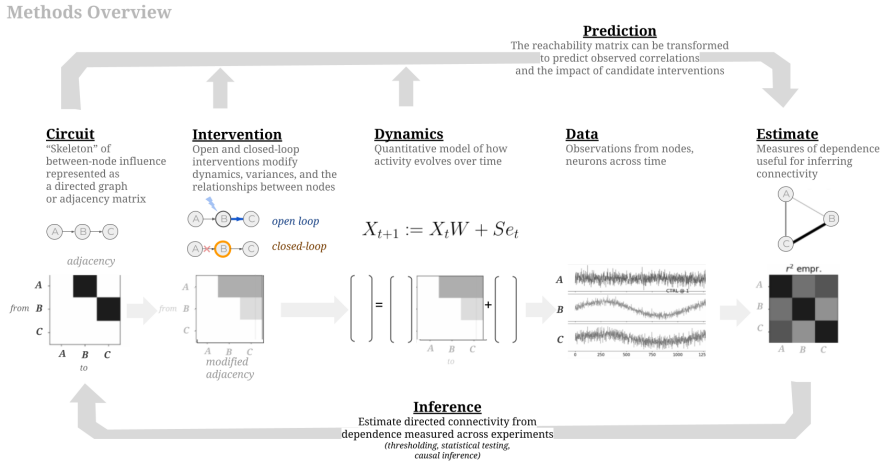
probably want - two circuits which look clearly different - ! but which have equivalent reachability - possibly with reciprocal connections - possssibly with common modulation

- do we need to reflect back from set of possible observations to consistent hypotheses?
  - mention markov equivalence classes explicitly?

- intuitive explanation using binary reachability rules
- *point to the rest of the paper as deepening and generalizing these ideas*
- *(example papers - Advancing functional connectivity research from association to causation, Combining multiple functional connectivity methods to improve causal inferences)*
- connect **graded reachability** to ID-SNR
  - IDSNR_{ij} measures the strength of signal related to the connection $i \to j$ relative to in the output of node j
  - for true, direct connections this quantity increasing means a (true positive) connection will be identified more easily (with high certainty, requiring less data)
  - for false or indirect connections, this quantity increasing means a false positive connection is more likely to be identified
  - as a result we want to maximize IDSNR for true links, and minimize it for false/indirect links

(see also `sketches_and_notation/walkthrough_EI_dissection.md`)

Theory / Prediction



> **Figure OVERVIEW:** ...

Predicting correlation structure (theory)

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

of each of the p nodes in the circuit, and $W \in \mathbb{R}^{p\times p}$ denote the matrix of connection strengths such that

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

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[10]

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

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.[11] 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.

[10] 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 $\Sigma = \mathbb{E}\left[(I - W^T)^{-1}EE^T(I - W^T)^{-1}\right] = (I - W^T)^{-1}\text{cov}(E)(I - W^T)^{-T} = (I - W^T)^{-1}\text{diag}(s)(I - W^T)^{-T}$.

[11] 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 "/section_content/methods0_simulations_interventions_estimates.md"

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

@ import "/section_content/results1_impact_of_intervention.md"

Discussion

see limitations_future_work.md

References

see pandoc pandoc-citations

Supplement