1. 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.

2. Introduction

2.1. 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 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 \cite{chicharro2012when}. *Closed-loop* interventions ...@Adam: short description of closed-loop in neuro, maybe drawing from text in this collapsable:

▶ Proposal text to draw from:

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 \cite{pearl2009causality} \cite{maathuis2016review} \cite{chis2011structural}, 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 \cite{wiener1956theory}. This concept 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 \cite{melssen1987detection}. 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 \cite{dean2016dangers}. 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 \cite{dean2016dangers}. 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 \cite{knox1981detection} \cite{salinas2001correlated} \cite{garofalo2009evaluation}.

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 \cite{schreiber2000measuring} \cite{barnett2009granger}. The *transfer entropy* $T_{X\to Y}(t) = I(Y_t \colon X_{<t} \mid Y_{<t})$ 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* $T_{X\to Y|Z}(t) = I(Y_t \colon X_{<t} \mid Y_{<t}, Z_{<t})$ conditions on the past of other variables to account for their potential 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 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 \cite{janzing2013quantifying} or an ability to perturb the system \cite{ay2008information} \cite{lizier2010differentiating}. 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., \cite{shorten2021estimating}). 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 \cite{wibral2014directed}.

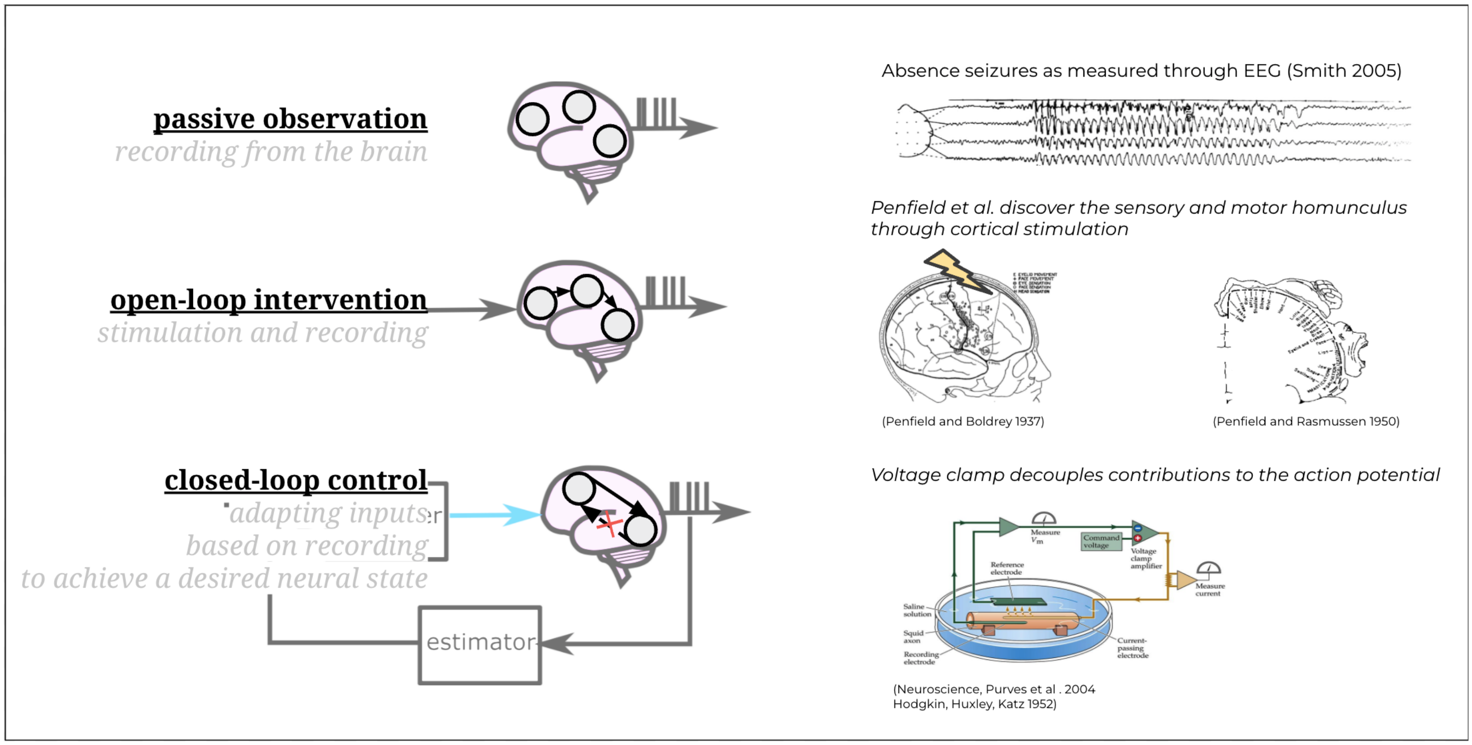
```
 [If we end up making the jump to IDTxl in our results: In our empirical results using transfer-entropy-based notions of directional in
```

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 \cite{garofalo2009evaluation}.

2.2. Interventions in neuroscience & causal inference

> **50% done:**

Data collected from experimental settings can be much more powerful than observational data alone. For example, passive observations of two correlated variables x and y do 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. Consistent with intuition from neuroscience literature, a rich theoretical literature has described the central role of interventions in inferring causal structure from data \cite{pearl2009causality, eberhardt2007interventions}.


passive observation
recording from the brain
open-loop intervention
stimulation and recording
closed-loop control
adapting inputs
based on recording
to achieve a desired neural state
estimator
Absence seizures as measured through EEG (Smith 2005)
Penfield et al. discover the sensory and motor homunculus through cortical stimulation
EYELID MOVEMENT
FACE MOVEMENT
EYE SENSATION
FACE SENSATION
HEAD SENSATION
(Penfield and Boldrey 1937)
(Penfield and Rasmussen 1950)
Voltage clamp decouples contributions to the action potential
Measure V_m
Command voltage
Voltage clamp amplifier
Reference electrode
Measure current
Saline solution
Squid axon
Recording electrode
Current-passing electrode
(Neuroscience, Purves et al . 2004
Hodgkin, Huxley, Katz 1952)

> **Figure INTRO:** Examples of the roles interventions have played in neuroscience. (A) *Passive observation* does not involve stimulating the brain. In this example, passive observational data is used to identify patients suffering from absence seizures. (B) *Open-loop stimulation* involves recording activity in the brain after perturbing a region with a known input signal. Using systematic *open-loop stimulation experiments*, Penfield uncovered the spatial organization of how senses and movement are mapped in the cortex \cite{penfield1937somatic} \cite{penfield1950cerebral}. (C) *Closed-loop control* uses feedback control to precisely specify activity in certain brain regions regardless of activity in other regions. Using closed-loop control, todo-Adam \citetodo-Adam.

The inferential power of interventions is also dependent on *where* stimulation is applied: interventions on some portions of a system may provide more information about the system's causal structure than interventions in other areas.[^precise] Interventions are also more valuable when they more effectively set the state of the system: "perfect" closed-loop control, which completely severs a node's activity from its inputs, are often more informative than "soft" interventions that only partially control a part of the system \cite{eberhardt2007interventions}.

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

Theoretical guarantees for the strength of these interventions -- algorithms designed to choose among them -- are often designed for simple models with strong assumptions.[2] However, they provide guidance that can that can help practitioners design experiments that provide as much scientific insight as possible.[3] Importantly, power of interventions is often independent of the algorithm used to infer causal connections, meaning that certain interventions can reveal portions of a circuit's causal structure that would be impossible for *any* algorithm to infer from only observational data (see, e.g., \cite{shanmugam2015learning}).

2.3. 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[4].

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[5]

$$X = XW + E,$$

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 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 t
```

transition from reachability to 2-circuit ID demo is now in background_id_demo.md

▶ ↪old reachability → ID demo text

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

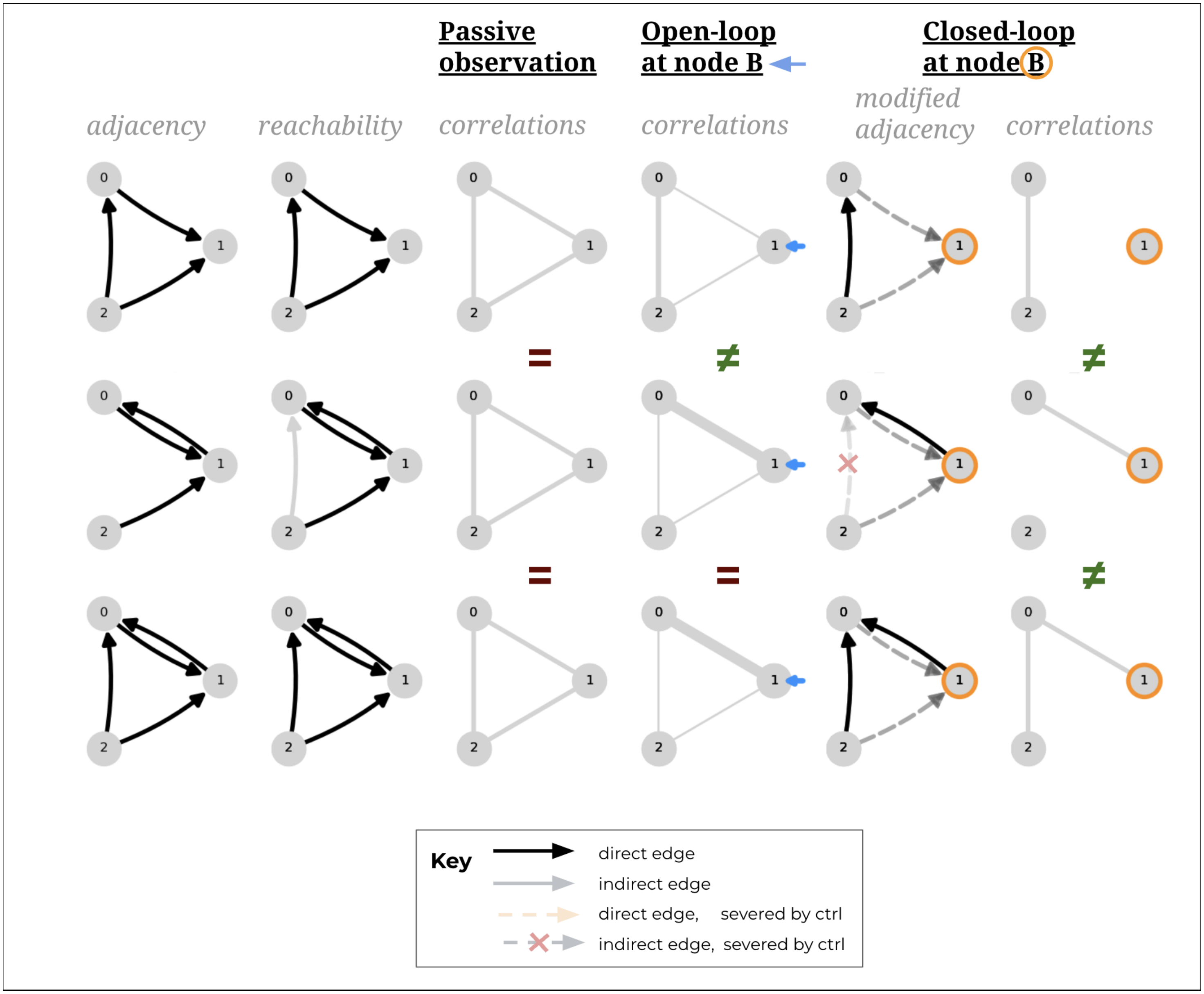


Figure DEMO *(box format)*: Applying CLINC to distinguish a pair of circuits

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 \ref{eq:adjacency-matrix}: 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 $2 \longrightarrow 1 \longrightarrow 0$ (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[6] for an experimentalist who performs only passive observation, but can be distinguished through stimulation.

Column 4 shows the impact on observed correlations of performing *open-loop* control on node 1. In hypothesis A, node 1 is not a driver of other nodes, so open-loop stimulation at this site will not increase the correlation between the signal observed at node 1 and other nodes. The path from node 1 to 0 in hypotheses B and C, meanwhile, causes the open-loop stimulation at node 1 to *increase* the observed correlation between nodes 1 and 0. An experimenter can thus distinguish between hypothesis A and the other two hypotheses by 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.

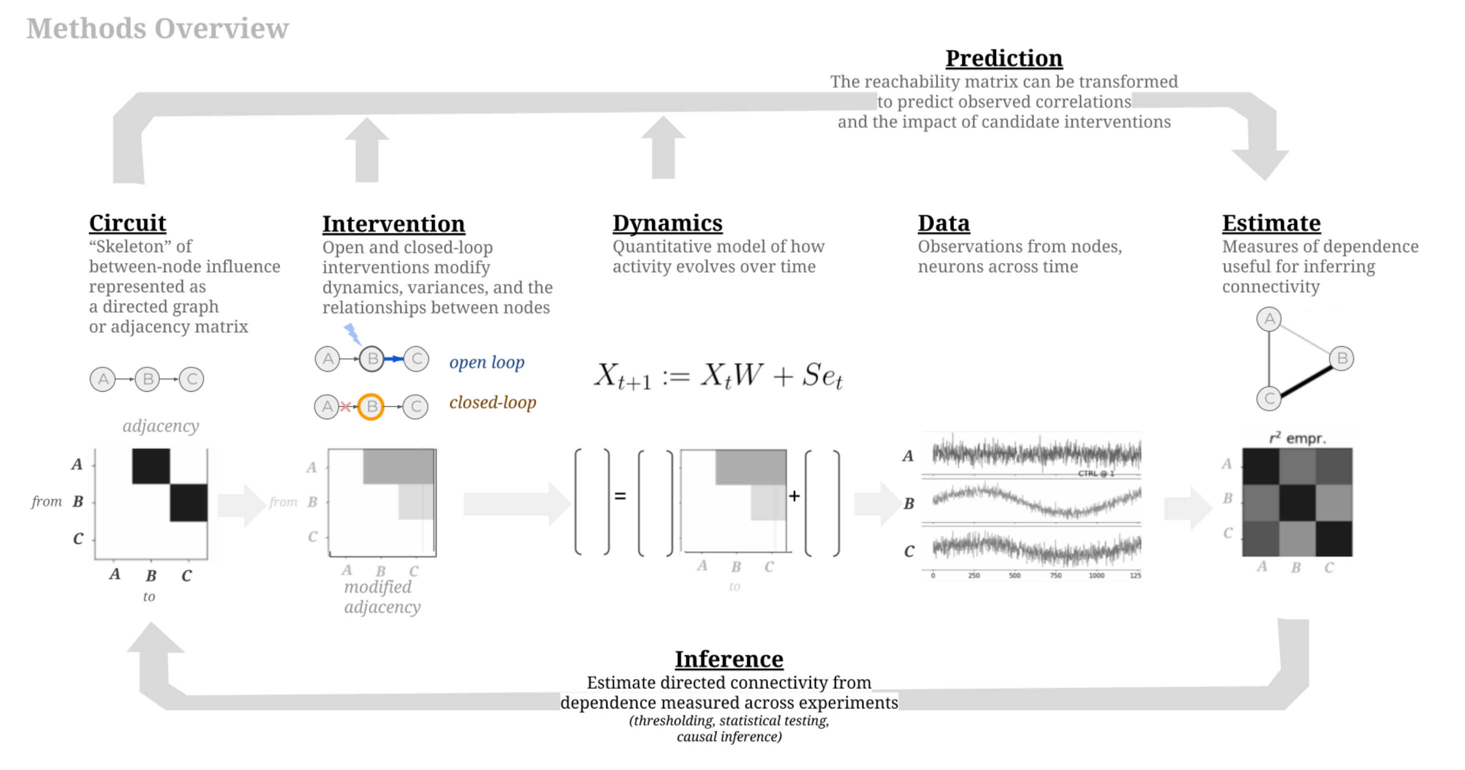
Closed-loop control (columns 5 and 6) can provide the experimenter with even more inferential power. Column 5 shows the resulting adjacency matrix when this closed-loop control is applied to node 1. In each hypothesis, the impact of this closed-loop control is to remove the impact of other nodes on node 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

▶ ↪2,3 circuit versions, straight from code
▶ ↪to do items
▶ ↪see also
▶ ↪more notes

3. Theory / Prediction



> **Figure OVERVIEW:** ...

3.1. 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[7]

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

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.[8] 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.

4. Simulation Methods

@ import "methods0_0_overview.md"

4.1. Modeling network structure and dynamics

70% 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.

4.1.1. Stochastic network dynamics

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[9].

Neuron dynamics:

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

4.1.2. Time-resolvable interactions

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

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

> correlation across positive and negative lags between two outputs

In the delay-resolvable domain, directionality of connections may be inferred even under passive observations by looking at temporal precedence - whether the past of one signal is more strongly correlated with future lags of another signal *(i.e. cross-correlation)*. In the contemporaneous domain, network influences act within the time of a single sample[10] so this temporal precedence clue is lost (although directionality can still be inferred in the presence of intervention).

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

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

▶ ↪concept figures

4.1.3. Code implementation

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

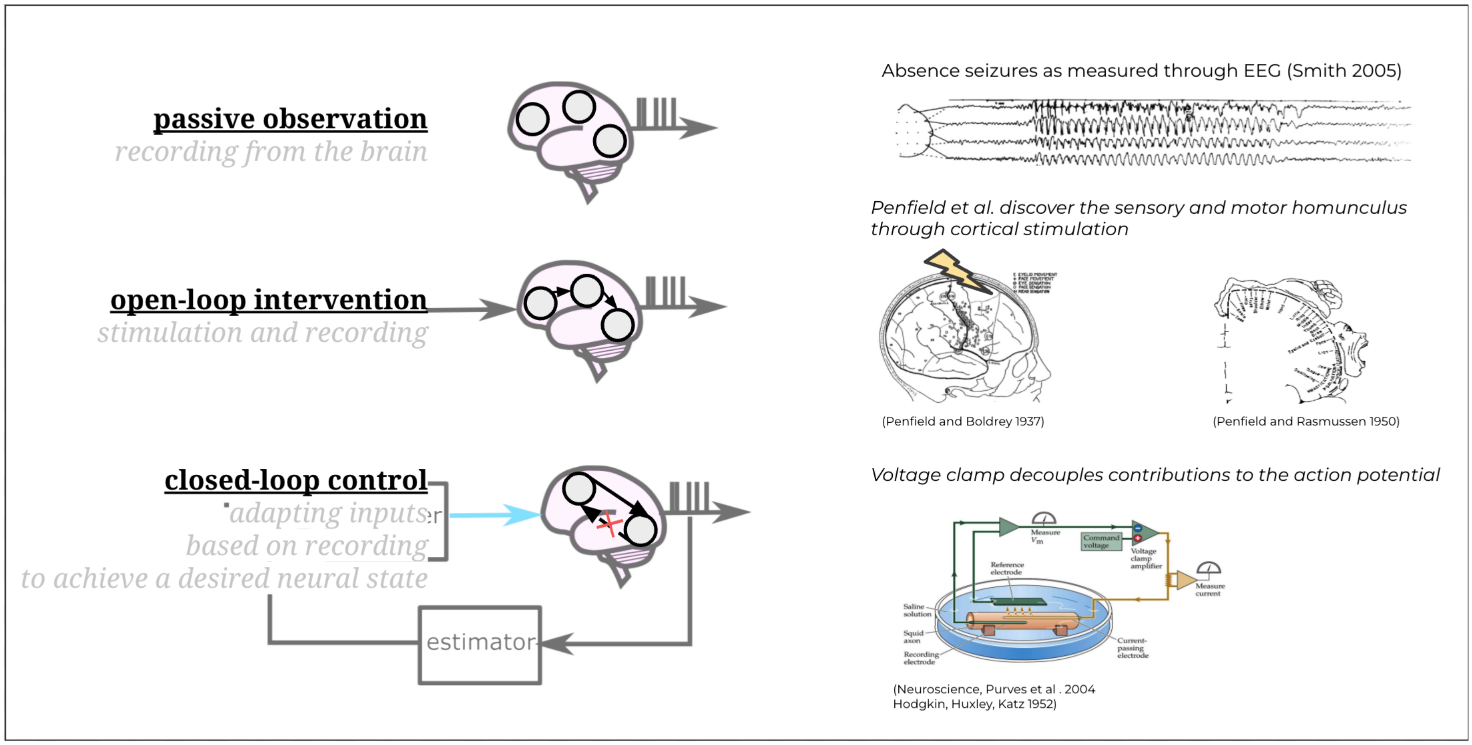
talk about parameter choices and ranges?

see _network_parameters_table.md for list of relevant parameters

4.2. Implementing interventions

70% done

assumed: effect of interventions on theory already address


passive observation
recording from the brain
open-loop intervention
stimulation and recording
closed-loop control
adapting inputs
based on recording
to achieve a desired neural state
estimator
Absence seizures as measured through EEG (Smith 2005)
Penfield et al. discover the sensory and motor homunculus through cortical stimulation
E EYELID MOVEMENT
F FACE MOVEMENT
EYE SENSATION
FACE SENSATION
HEAD SENSATION
(Penfield and Boldrey 1937)
(Penfield and Rasmussen 1950)
Voltage clamp decouples contributions to the action potential
Measure V_m
Command voltage
Voltage clamp amplifier
Reference electrode
Measure current
Saline solution
Squid axon
Recording electrode
Current-passing electrode
(Neuroscience, Purves et al . 2004
Hodgkin, Huxley, Katz 1952)

To study the effect of various interventions we simulated inputs to nodes in a network. In the **passive setting**, nodes receive additive drive from *private* Gaussian noise sources common to all neurons within a node, but independent across nodes. The variance of this noise is specified by $\sigma_m\sqrt{\tau_m}$.[11]

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

To emulate **open-loop intervention** we simulated current injection from an external source. This is intended to represent experiments involving stimulation from microelectrodes or optogenetics *(albeit simplifying away any impact of actuator dynamics)*. By default, open-loop intervention is specified as white noise sampled at each timestep from Gaussian distribution with mean and variance $\mu_{intv.}$ and $\sigma^2_{intv.}$[12]

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

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

$$X_k = f(\sigma^2_m, \sigma^2_{intv.})$$

```
per−node indexing needs resolving here also
```

Ideal **closed-loop control** is able to overwrite the output of a node, setting it precisely to the specified target.

```
making up notation as I go here, needs tightening up:
```

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

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

$$X_k | CL_k = f(\sigma^2_{intv.}) \perp \sigma^2_m$$

In practice, near-ideal control is only possible with very fast measurement and computation relative to the network's intrinsic dynamics, such as in the case of dynamic clamp[13]. To demonstrate a broader class of closed-loop interventions (such as those achievable with extracellular recording and stimulation), imperfect "partial" control is simulated by linearly interpolating the output of each node between the target T and the uncontrolled output based on a control effectiveness parameter γ

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

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

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

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

Because the instantiation of noise in the network will be different from trial to trial, this "replay" stimulus will no longer adapt sample-by-sample (therefore it should be considered open-loop) and the node's output cannot be expected to match the target precisely, however the statistics of externally applied inputs will be the same. In effect, the comparison between closed-loop and open-loop replay conditions reveals the specific effect of feedback intervention while controlling for any confounds from input statistics.

4.3. Extracting circuit estimates

10% done

> *refer to methods overview figure*

While a broad range of techniques[14] exist for inferring functional relationships from observational data, `(for the majority of this work)` we choose to focus on simple bivariate correlation as a measure of dependence in the linear-Gaussian network. The impact of intervention on this metric is analytically tractable *(see methods1_predicting_correlation.md)*, and can be thought of as a prototype[15] for more sophisticated measures of dependence such as time-lagged cross-correlations, bivariate and multivariate transfer entropy.

We implement a naive comparison strategy to estimate the circuit adjacency from emprical correlations; Thresholded empirical correlation matrices are compared to correlation matrices predicted from each circuit in a hypothesis set. Any hypothesized cirucits which are predicted to have a similar correlation structure as is observed (i.e. corr. mats equal after thresholding) are marked as "plausible circuits."[16] If only one circuit amongst the hypothesis set is a plausible match, this is considered to be the estimated circuit. The threshold for "binarizing" the empirical correlation matrix is treated as a hyperparameter to be swept at the time of analysis.[17]

4.4. Information-theoretic measures of hypothesis ambiguity

> **10% done**

see _steps_of_inference.md for entropy writeup

5. Results

> **overall, 60% done**

5.1. Impact of intervention on estimation performance

5.1.1. 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[18] to illustrate the impact of intervention. 🚧 `most of the story in the figure caption for now` 🚧

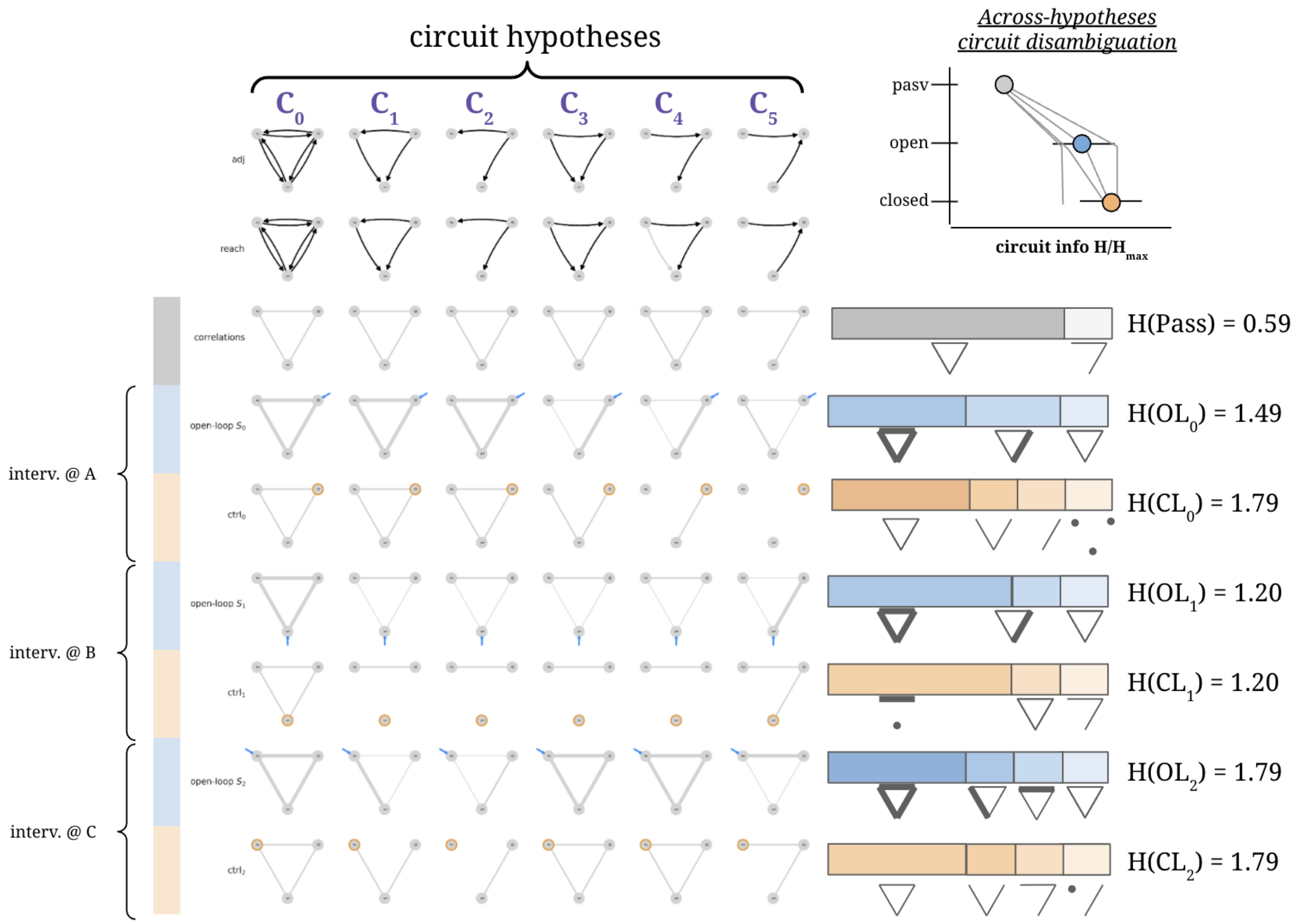


Figure DISAMBIG: Interventions narrow the set of hypotheses consistent with observed correlations

source: google drawing

(A) Directed adjacency matrices represent the true and hypothesized causal circuit structure

(B) Directed reachability matrices represent the direct *(black)* and indirect *(grey)* influences in a network. Notably, different adjacency matrices can have equivalent reachability matrices making distinguishing between similar causal structures difficult, even with open-loop control.

(C) Correlations between pairs of nodes. Under passive observation, the direction of influence is difficult to ascertain. In densely connected networks, many distinct ground-truth causal structures result in similar "all correlated with all" patterns providing little information about the true structure.

(D-F) The impact of open-loop intervention at each of the nodes in the network is illustrated by modifications to the passive correlation pattern. Thick orange[19] edges denote correlations which increase above their baseline value with high variance open-loop input. Thin blue[19:1] 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**[20] strongly determines the inference power of your experiment.
> **secondary point:** having (binary) prediction helps capture this relationship

Quantitative impact of closed-loop - 70% done

5.1.2. 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[21] 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[22]

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

Applying closed-loop to a linear gaussian circuit:

$$\begin{aligned}\mathbb{V}_i(C|S=\text{closed},\sigma_S^2) &= \sigma_S^2 \qquad (1)\\ \mathbb{V}_i(C|S=\text{closed},\sigma_S^2) &\perp \mathbb{V}_i(C) \qquad (2)\end{aligned}$$

▶ ↪ Firing rates couple mean and variance

▶ ↪ Notes on imperfect control

reference figvar to empricially show this bidirectional control of output variance?

5.1.2.1. 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[24] `[why do we want to?...]`

One application of this increased flexibility [...] is to increase correlations associated with pairs of directly correlated nodes, while decreasing spurious correlations associated with pairs of nodes without a direct connection (but perhaps are influenced by a common input, or are connected only indirectly). This manipulation may bring the observed pattern of correlations

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

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

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

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

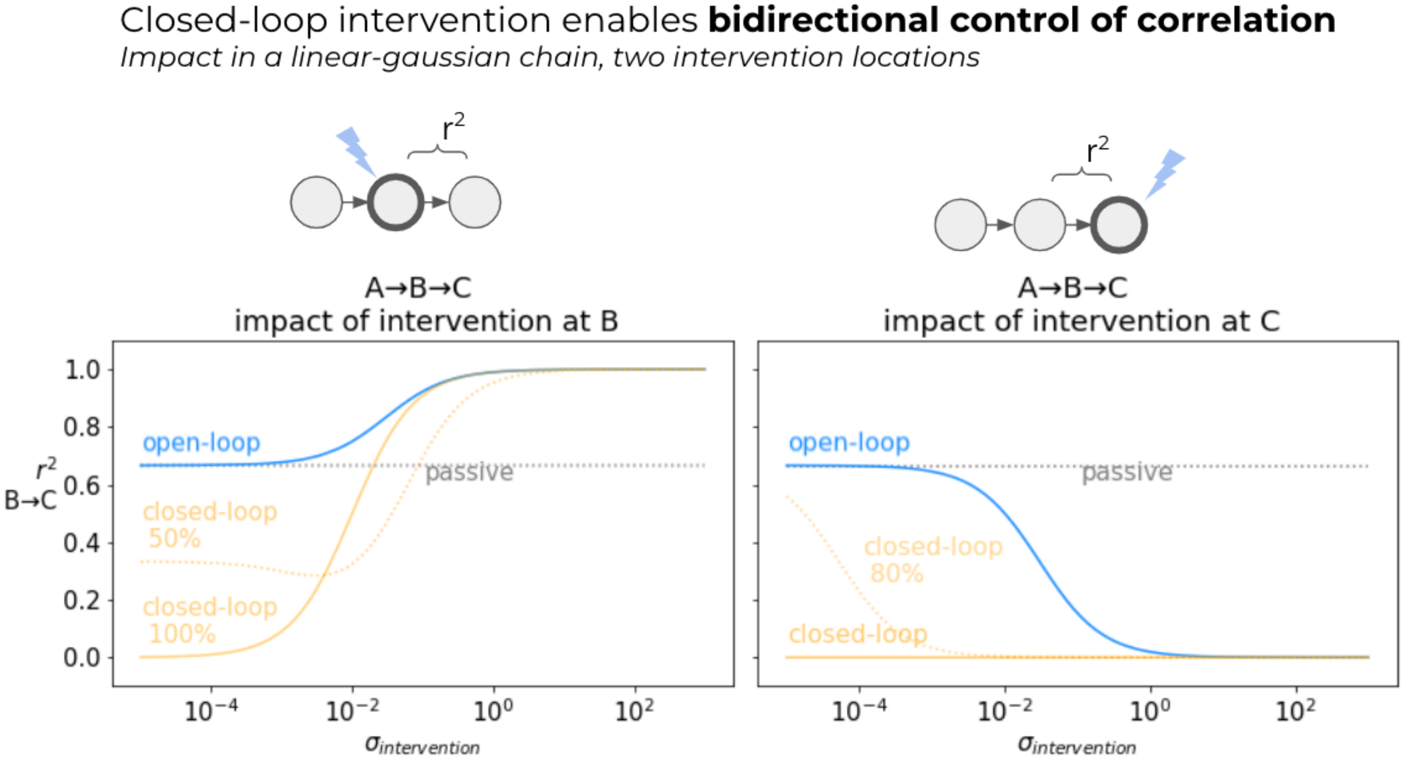
Second, interventions affecting only the downstream node (lower right, fig. variance) of a true connection introduce variance which is independent of the connection A→B, decreasing the observed correlation.

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

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

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

Notably, the impact of an intervention which is a "common cause" for both nodes depends on the relative weighted reachability between the source and each of the nodes. Correlations induced by a common cause are maximized when the input to each node is equal, that is $\widetilde{W}_{S_k \to i} \approx \widetilde{W}_{S_k \to j}$ (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 ($\frac{dR}{dS_k} < 0$)[25]



> 🚧 (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 σ_S^2. Closed-loop interventions *(orange)* consist of feedback control with an independent gaussian target with a range of variances. *Incomplete closed-loop interventions result in node outputs which are a mix of the control target and network-driven activity*. Connections from sources to nodes are colored by their impact on correlations between A and B; green denotes $dR/dS > 0$, red denotes $dR/dS < 0$.
> **(lower left)** Intervention "upstream" of the connection A→B increases the correlation $r^2(A, B)$.
> **(lower right)** Intervention at the terminal of the connection A→B decreases the correlation $r^2(A, B)$ by adding connection-independent noise.
> **(upper left)** Intervention with shared inputs to both nodes generally increases $r^2(A, B)$, *(even without A→B, see supplement)*.
> **(upper right)** The impact of shared interventions depends on relative weighted reachability $\mathrm{Reach}(S_k \to A)/\mathrm{Reach}(S_k \to B)$, with highest correlations when these terms are matched (see *)
> Closed-loop interventions *(orange)* generally result in larger changes in correlation across σ_S^2 than the equivalent open-loop intervention. Closed-loop control at B effectively lesions the connection A→B, resulting in near-zero correlation.
> [26]

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

see also results1B_data_efficiency_and_bias.md

6. Discussion

see limitations_future_work.md

7. References

see pandoc pandoc-citations

8. Supplement

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 ↩
2. These assumptions are typically on properties such as the types of functional relationships that exist in circuits, the visibility and structure of confounding relationships, and noise statistics. ↩
3. if citations needed here, could start by looking for a good high-level reference in either \cite{ghassami2018budgeted} or \cite{yang2018characterizing}. (Both of these papers are pretty technical, so likely wouln't be great citations on their own.) ↩
4. 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. ↩
5. 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 ↩

6. saying "difficult to distinguish" instead of "indistinguishable" here since the magnitudes of the correlations could also be informative with different assumptions ↩
7. 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 $\Sigma = \mathrm{cov}(X) = \mathbb{E}\left[XX^T\right]$ can then be written as $\Sigma = \mathbb{E}\left[(I - W^T)^{-1} E E^T (I - W^T)^{-1}\right] = (I - W^T)^{-1} \mathrm{cov}(E)(I - W^T)^{-T} = (I - W^T)^{-1} \mathrm{diag}(s)(I - W^T)^{-T}$. ↩
8. 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. ↩
9. However, depending on overall firing rates and population sizes, this sparse spike-based transmission can be coarse-grained to a firing-rate-based model. ↩
10. the effective Δ_{sample} would be broadened in the presence of jitter in connection delay, measurement noise, or temporal smoothing applied post-hoc, leading ↩
11. need to triple check indexing w.r.t. nodes, neurons ↩
12. need to resolve differences in implementation between contemporaneous and voltage simulation cases ↩
13. NEED dynamic clamp refs - http://www.scholarpedia.org/article/Dynamic_clamp ↩
14. *inference techniques mentioned in the intro...* ↩
15. what does "prototype" mean here? something like MI and corr are equivalent in the linear-Gaussian case, ... ↩
16. TODO? formalize notation for this ↩
17. not sure how important this is. would prefer to set this threshold at some ad-hoc value since we're sweeping other properties. But a more in-depth analysis could look at a receiver-operator curve with respect to this threshold ↩
18. 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. ↩
19. 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 ↩ ↩
20. Figure VAR shows this pretty well, perhaps sink this section until after discussing categorical and quantitative? ↩
21. below the level set by added, independent/"private" sources ↩
22. notably, this is part of the definition of open-loop intervention ↩
23. 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. ↩
24. 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) ↩
25. 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. ↩
26. 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. ↩