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 sim`

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

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 (Chicharro and Ledberg 2012). *Closed-loop* interventions […]

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 (Pearl 2009; Maathuis and Nandy 2016; Chis, Banga, and Balsa-Canto 2011), 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`[2] 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 (Wiener 1956). This concept has traditionally been operationalized with vector autoregressive models (Granger 1969); the requirement that *all* potentially causative variables be considered makes these notions of dependence susceptible to unobserved confounders (Runge 2018).

Our work initially focuses on measures of directional interaction that are based on lagged correlations[3] (Melssen and Epping 1987). 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 (Dean and Dunsmuir 2016). 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 (Dean and Dunsmuir 2016). 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 (Knox 1981; Salinas and Sejnowski 2001; Garofalo et al. 2009).

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 (Schreiber 2000; Barnett, Barrett, and Seth 2009). 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 $Y_{<t}$)

(Bossomaier et al. 2016). Using transfer entropy as a measure of causal interaction requires accounting for potential confounding variables; the *conditional transfer entropy* \(T_{X \to Y \mid Z}(t) = I(Y_t \colon X_{<t} \mid Y_{<t}, Z_{<t})\) conditions on the past of other variables to account for their potential confounding influence `Sec.~4.2.3` (Bossomaier et al. 2016). 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 (Janzing et al. 2013) or an ability to perturb the system (Ay and Polani 2008; Lizier and Prokopenko 2010). 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., (Shorten, Spinney, and Lizier 2021)). 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 (Wibral, Vicente, and Lizier 2014).
`[If we end up making the jump to IDTxl in our results: In our empirical results`
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 (Garofalo et al. 2009).

Interventions in neuroscience & causal inference

Data collected from experimental settings can provide more inferential power than observational data alone. For example, consider an experimentalist who is considering multiple causal hypotheses for two nodes under study, \(x\)

and \(y\): the hypothesis that \(x\) is driving \(y\), the hypothesis that \(y\) is driving \(x\), or the hypothesis that the two variables are being independently driven by a hidden confounder. Observational data revealing that \(x\) and \(y\) produce correlated time-series data is equally consistent with each of these three causal hypotheses, providing the experimentalist with no inferential power. Experimentally manipulating \(x\) and observing the output of \(y\), however, allows the scientist to begin to establish which 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 (Pearl 2009; Eberhardt and Scheines 2007).

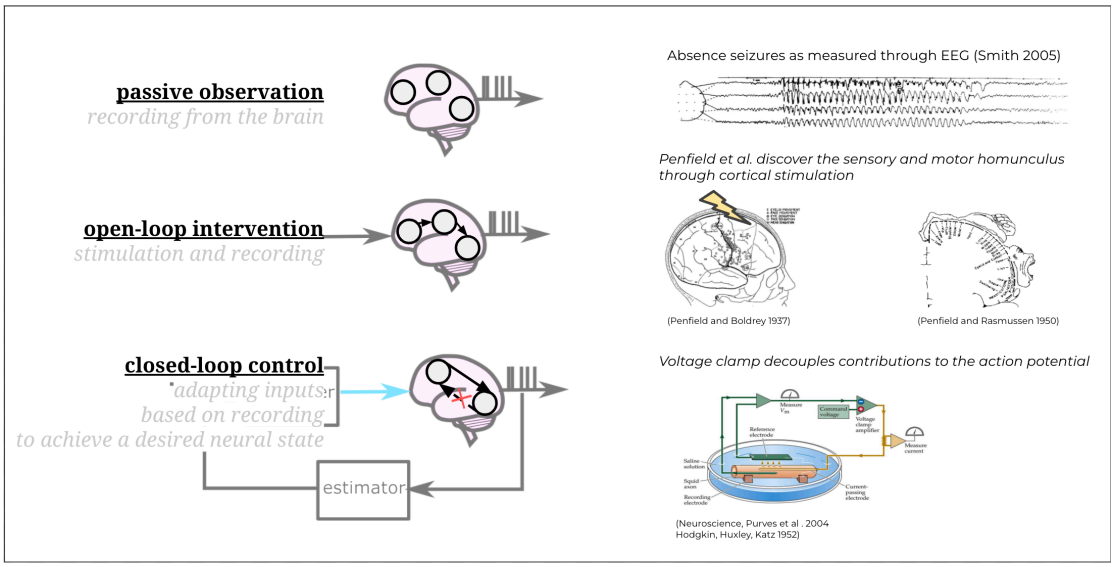


Figure INTRO: Examples of the roles interventions have played in neuroscience. (A) *Passive observation* does not involve stimulating the brain. In this example, passive observational data is used to identify patients suffering from absence seizures. (B) *Open-loop stimulation* involves recording activity 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 (W. Penfield and Boldrey 1937 ; Wilder Penfield and Rasmussen 1950). (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,

<...> .[4]

The inferential power of interventions is depends 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. And 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 (Eberhardt and Scheines 2007).

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.

Although algorithms designed to choose optimal interventions are often designed for simple models with strong assumptions,[5] they provide intuition that can aid practitioners seeking to design real-world experiments that provide as much scientific insight as possible.[6] Importantly, the informativeness 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 [Das and Fiete (2020)][7]. We similarly expect the results we demonstrate in this paper to both inform experimentalists and open avenues for further research.

Representations & reachability (minimal, dupe)

```
consider:
@ import "/section_content/representation_reach.md"
@ import "/section_content/background_id_demo.md"
```

Results

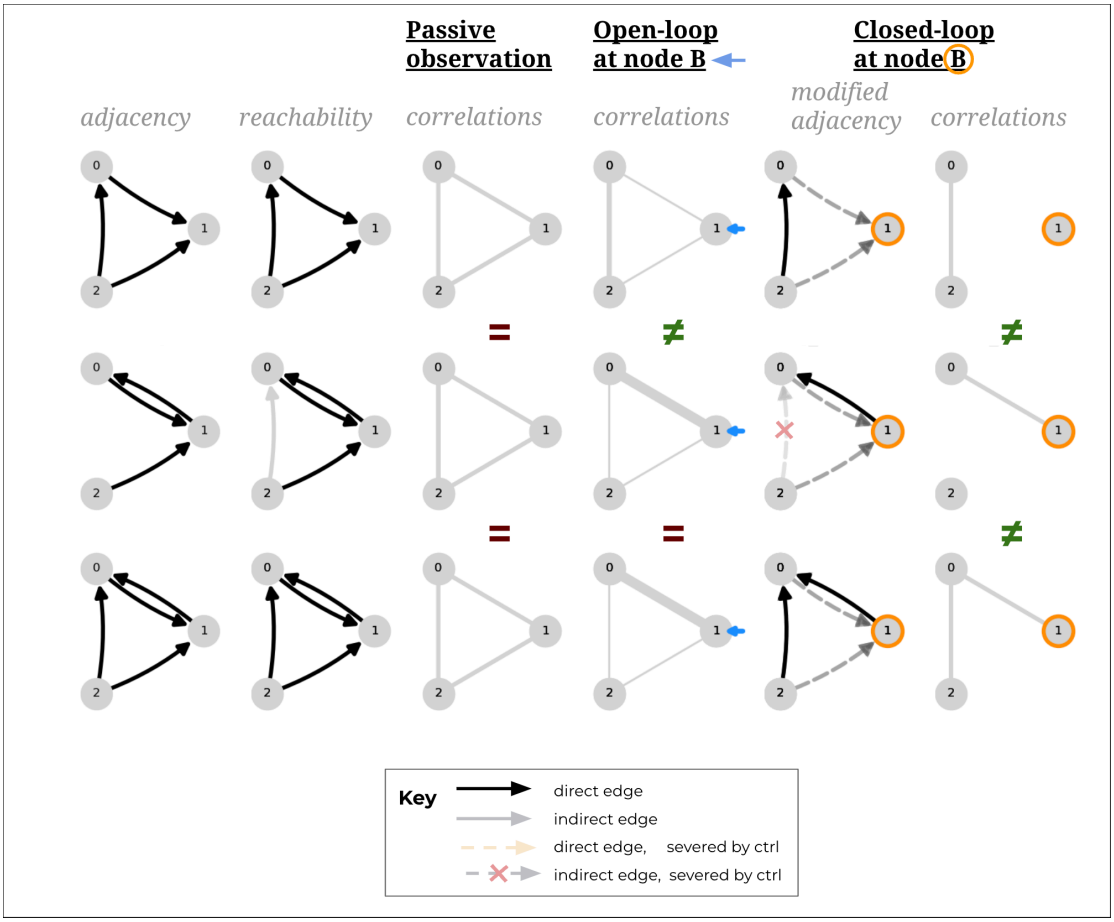


Figure DEMO *(box format)*: Applying CLINC to distinguish a pair of circuits

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 \(\to\) 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[8] 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.

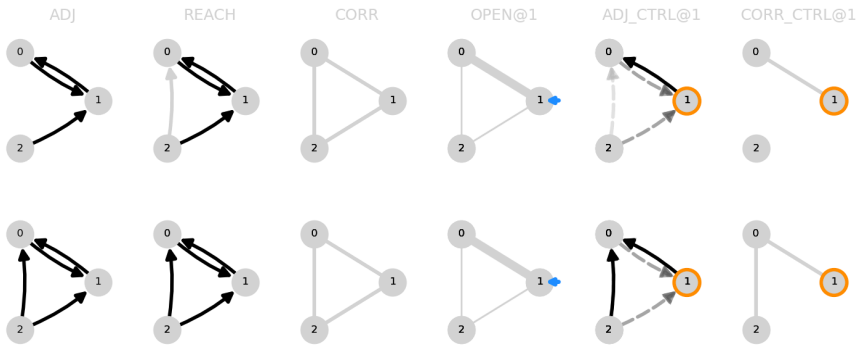
Closed-loop control (columns 5 and 6) can provide the experimenter with even more inferential power. Column 5 shows the resulting adjacency matrix when this closed-loop control is applied to node 1. In each hypothesis, the impact of this closed-loop control is to remove the impact of other nodes on node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

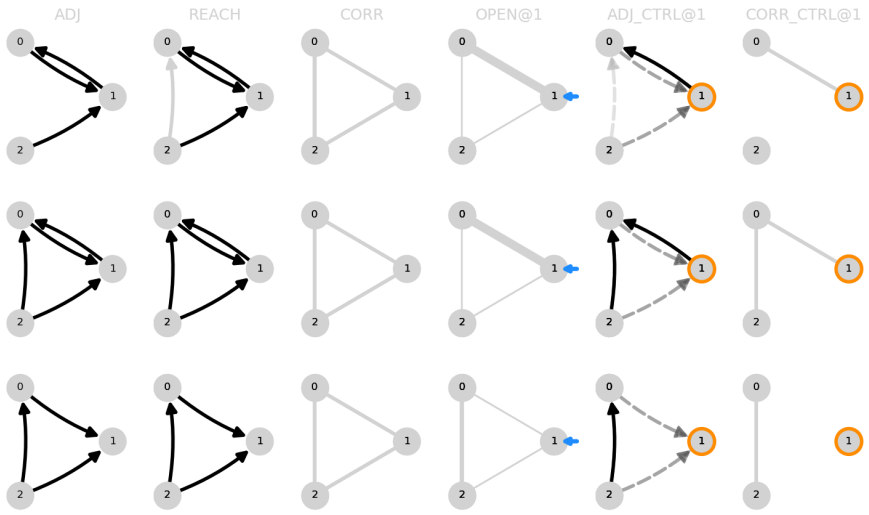
> control at node 1 thus results in a different observed correlation structure in each of the three circuit hypotheses (column 6). This means that the experimenter can therefore distinguish between these circuit hypotheses by applying closed-loop control – a task not possible with passive observation or open-loop control.

↪ figure to do items for “Adam-to-Do” (2022)

- [ ] TODO: overall this needs to be cut from the caption and filtered into the text body
- [ ] “Adam-to-Do” (2022) - change labels at top from “B” to “1”
- [ ] “Adam-to-Do” (2022) - add (A) (B) (C) labels to each row
- [ ] “Adam-to-Do” (2022) - in legend, change in/direct “edge” to in/direct “connection”
- [ ] “Adam-to-Do” (2022) - in legend, orange dashed arrow to dark gray

↪2,3 circuit versions, straight from code





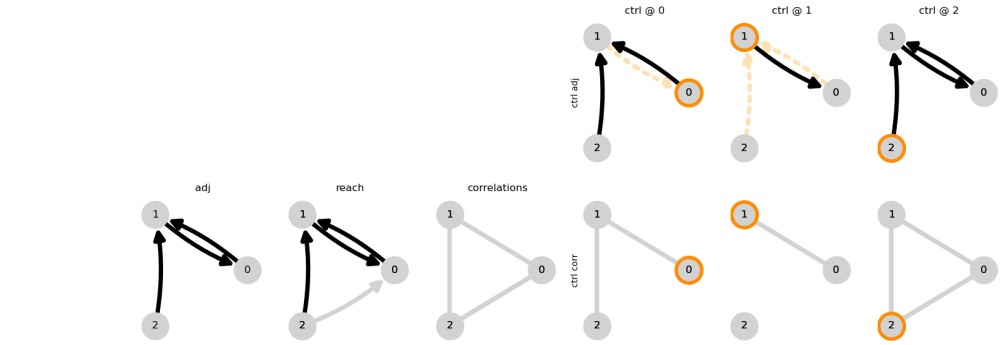
> 3 circuit walkthrough, walkthrough will all intervention locations might be appropriate for the supplement

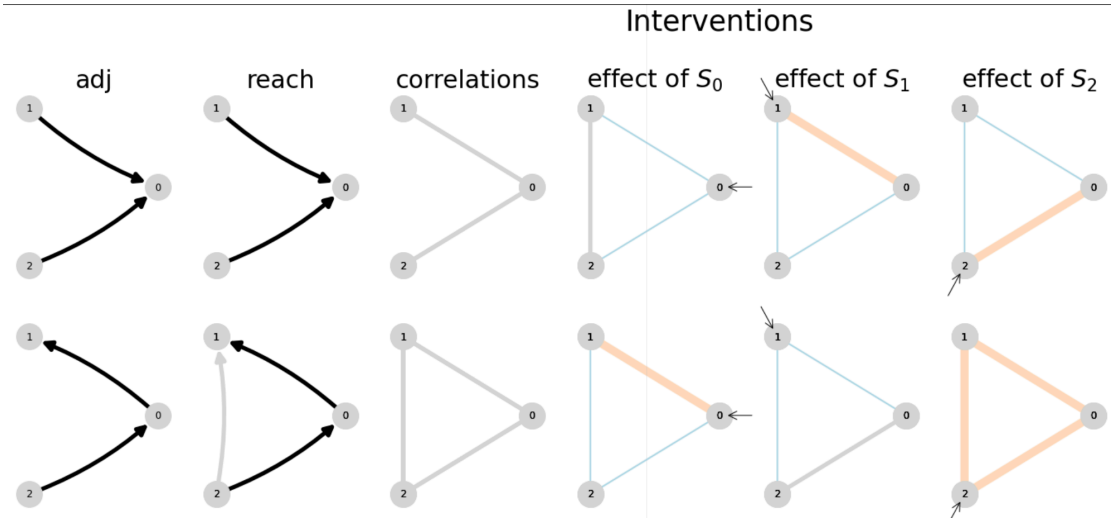
↪to do items

- [ ] 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.simple_paths.all_simple_paths.html#networkx.algorithms.simple_paths.all_simple_paths

↪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

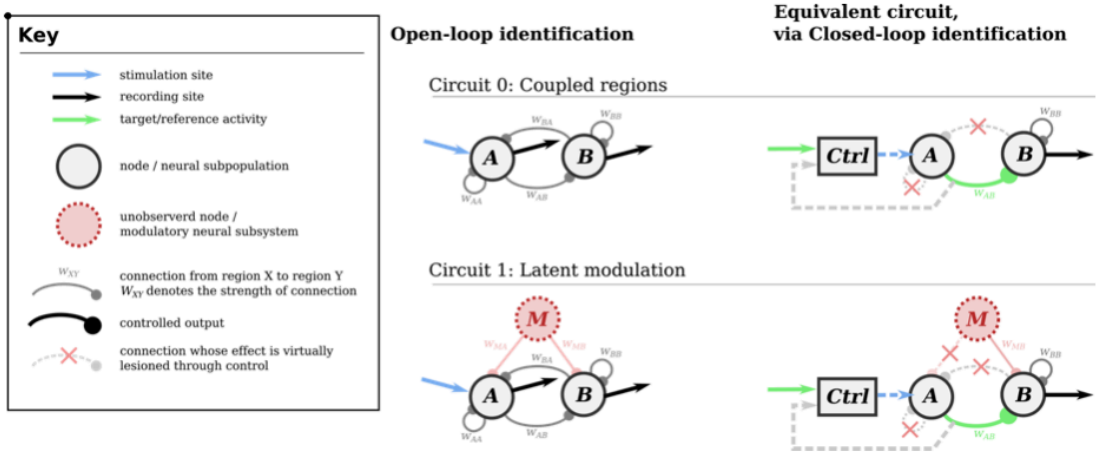


Figure 11: Closed-loop control compensates for inputs to a node in simple circuits: *The left column shows a simple circuit and recording and stimulation sites for an open-loop experiment. The right column shows the functional circuit which results from closed-loop control of the output of region A.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

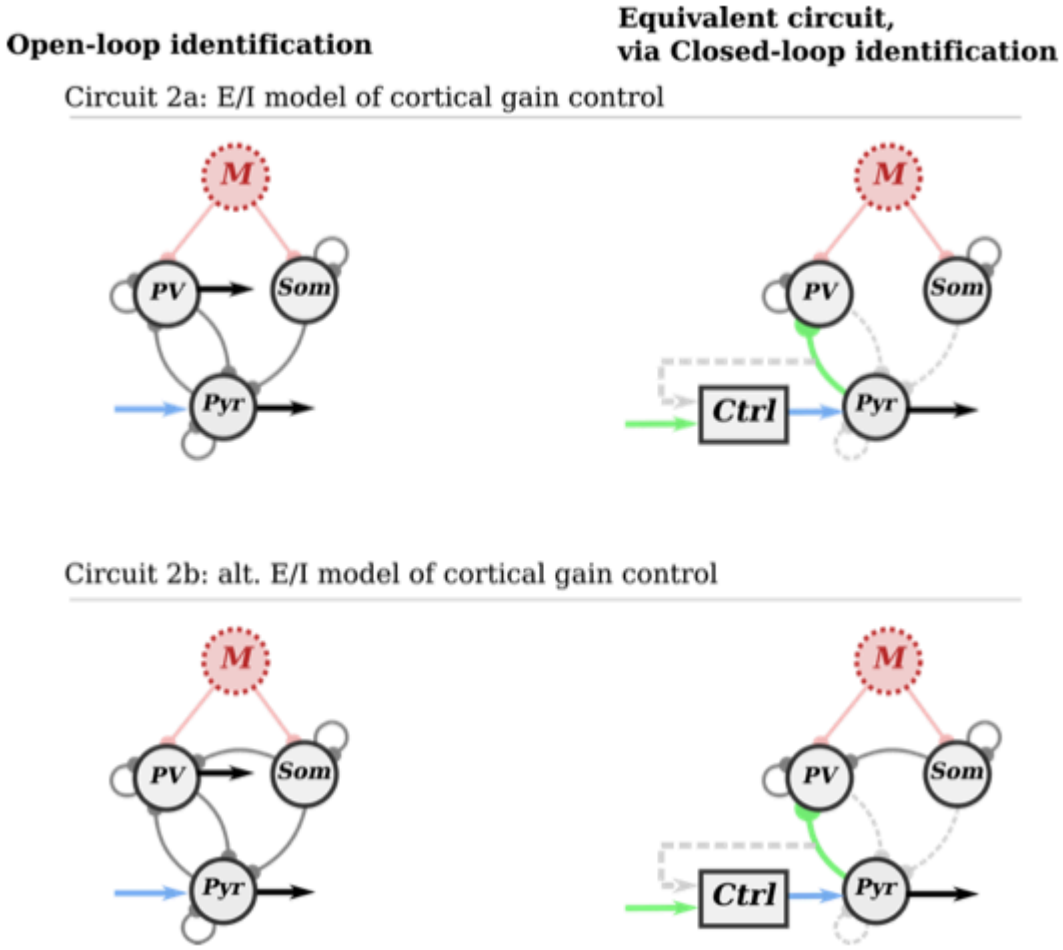


Figure 12: Closed-loop control allows for two circuit hypotheses to be distinguished. Two hypothesized circuits for the relationships between pyramidal (Pyr, excitatory), parvalbumin-positive (PV, inhibitory), and somatostain-expressing (Som, inhibitory) cells are shown in the two rows. Dashed lines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

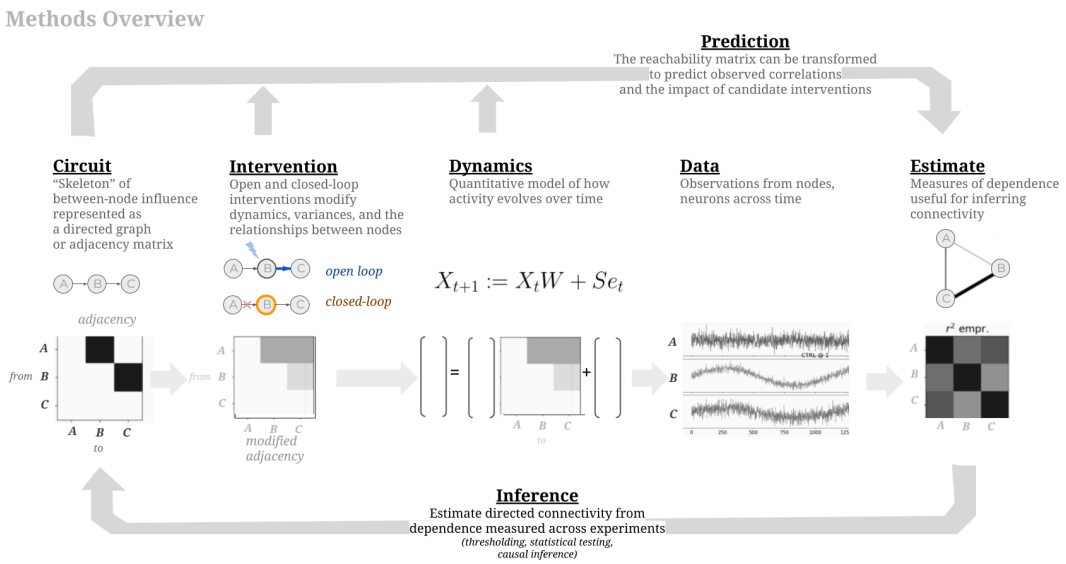
`reference extended methods`

```
extract minimum from:
@ import
"/section_content/methods_simulations.md"
```

- `reference extended methods`

```
extract minimum from:
@ import "/section_content/methods_simulations.md"
```

Steps of inference - *overview of CLINC approach* (+)



> **Figure OVERVIEW:** …

> **Theme B.** Experiments for circuit inference can be thought of as **narrowing the set of plausible explanations**, refining a hypotheses space[9]

We envision the structure of an experiment[10] to include the following broad stages:

1. First, explicitly **enumerate the set of hypothesized circuits.** Hypotheses about the structure of the circuit would be based on

multiple sources of information including prior recordings, anatomical constraints revealed by `experiments where you look at the fiber bundles connecting regions`, or commonly observed connectivity patterns in other systems `[🚧 add other sources of priors for circuit hypotheses]`[^bonus_causal][^more_assumptions] These hypotheses should be expressed as a set of circuits (adjacency matrices) each with a probability representing the prior belief about the relative likelihood of these options. This hypothesis set can be thought of as a space of possible explanations for the observed data so far, which will be narrowed down through further intervention, observation, and inference. (Fig.DISAMBIG top row)

2. Second, *in silico*, **forecast patterns of correlation** which could result from applying candidate interventions. Most algorithms[11] for circuit inference quantify and threshold measures of dependence between pairs of nodes. Correlations are often used to measure the linear component of dependence between outputs of two nodes, although the approach described here should generalize to other nonlinear measures of dependence such as mutual information. As such, the observed pattern of dependence (correlations) in a given experiment summarizes the input to an inference procedure to recover an estimated circuit.

   A detailed forecast of the observed outputs could be achieved by simulating biophysical networks across candidate interventions and hypothesized ground-truth circuits. However, for large networks or large hypothesis sets this may be expensive to compute. Instead, for the sake of rapid iteration in designing interventions, we propose using the reachability representation of a linear (linearized) network to succinctly and efficiently predict the observed correlations[12] across nodes[^node_repr]. The methods described in `[ref. prediction methods]` allow us to anticipate how open and closed-loop interventions across nodes in the network might increase, decrease, or sever dependencies between node outputs.

3. `{Survey / analyze / compare / summarize}` `{diversity / equivalence / distinguishability of}` patterns of correlation across each hypothesized circuit. A useful experiment (intervention) is one which produces highly distinct outcomes when applied to each of the hypothesized circuits, while an experiment which produces the same outcome across all hypothesized circuits would be redundant. Before collecting experimental data we do not know the ground-truth circuit with certainty, therefore it is useful to understand the range of possible observed patterns of dependence. To distill this range of possibilities to a make a decision about which intervention to apply, it is also useful to summarize the expected information we would gain about circuit identity across the range of hypotheses. (across columns of Fig.DISAMBIG) >-*Here we generalize across specific values of synaptic weights and divide observed patterns into categories: increased correlation, decreased correlation, no correlation.*

🚧 `Entropy as a measure of information about circuit hypotheses`

`@ import "/section_content/methods_entropy.md"`

`select intervention – (is this its own step, or the last part of step 3)`

Here, we describe a “greedy” approach for choosing an effective single-node intervention, but extending the approach above to predict joint entropy would allow a joint or sequential experimental design which would be optimal over multiple interventions. >- possible interventions consist of open-loop and closed-loop stim at each of N nodes > - but more constraints on the set of interventions can easily be incorporated at this stage

For selecting the first intervention type and location, we propose choosing the intervention which results in the maximum expected circuit information, that is: \[S_i^* = \underset{i}{\arg\max}\,H(C|S_i)\][13]

🚧 4. Apply intervention and collect data Using entropy as a metric to select a useful intervention, the next step is to conduct that interventional experiment, in-vivo or in a detailed simulation. Such an experiment may reveal outputs

patterns not fully captured by the linearized reachability representation.

`[extract correlations ...]` [14].

5. **Given the observed dependency pattern, form a posterior belief over hypotheses** `[🚧 transition text]`

`@ import "/section_content/methods_entropy_selection.md"`

Impact of intervention on estimation performance

(predicting) impact of intervention on pairwise dependence (3.1?, 5.1?)

Representations & reachability

`extract minimum from:`
`@ import "/section_content/representation_reach.md"`

Predicting correlation structure (theory)

```
extract minimum from:
@ import "/section_content/methods_predicting_correlation.md"
```

```
extract minimum from:
@ import "/section_content/methods_interventions.md"
@ import "/section_content/results_impact_of_intervention.md"
```

Intervening provides categorical improvements in inference power beyond passive observation

In the previous sections, we established how open-loop interventions modify observed pairwise correlations, and how closed-loop interventions modify a circuit's functional connectivity. Figure `ID-DEMO` demonstrated a simple example of how removing connections in a circuit can sometimes reveal

more distinct patterns of dependence, and distinguish hypotheses which are indistinguishable through passive observation and open-loop control. Here, we systematize this approach to choose an appropriate intervention to narrow down a hypothesis set. The following sections will address how to evaluate the relative effectiveness of a particular intervention. Multiple intervention types and locations are compared for a larger circuit hypothesis set to build towards general principles for where and how to intervene.

While the ground truth connectivity is rarely available during experiments, it is valuable to explicitly lay out our prior hypothesis in the form of a directed graph or adjacency matrix. Panel A of `Fig. DISAMBIG` shows the adjacency and reachability of 6 candidate circuit hypotheses. `Row Ba` illustrates the presence of pairwise correlation for each hypotheses under passive observation. While the magnitudes of correlation will depend on particular values of system parameters, here we focus on only the presence or absence of a significant correlation between two nodes, as well as whether correlations increase or decrease from their baseline. In this way, we build towards an understanding of the categorical impact of intervention on observed pairwise dependence, which should be general across particular parameter values or algorithms for circuit inference. *(More concrete, quantitative effects will be explored in the next section).*

The set of patterns of pairwise dependences across the hypothesis set form an “intervention-specific fingerprint” (i.e. a single row of `Fig. DISAMBIG`). This fingerprint summarizes the outcomes of a particular experiment with intervention, and therefore shows which hypotheses are observationally equivalent under this observation. If this fingerprint contains many examples of the same pattern (such as the all-to-all correlation pattern seen under passive observation, `row Ba`), many different circuits correspond to the same observation, and that experiment contributes low information to distinguish between hypotheses. On the other hand, a maximally informative experiment would result in unique observations corresponding to each hypothesis. Observations from such an experiment would be sufficient to narrow the inferred circuit down to a single hypotheses.

To quantify this hypothesis ambiguity based on the diversity of a set of

possible outcomes, we compute the Shannon entropy over the distribution of patterns (See Methods entropy). Because our hypotheses set contains circuits with relatively dense connectivity, 5 of the 6 hypotheses result in all-to-all correlations, with the final hypothesis displaying a unique V-shaped pattern of correlation (A~B, and A~C, `row Ba`). The entropy of this distribution is 0.65 bits. To interpret this entropy value, it is useful to understand the maximum achievable entropy, which is simply the logarithm of the number of hypotheses. In this case, \(H_{max} = \log_2(6)\approx 2.58 \text{bits}\), which indicates the information gained from passive observation is 25% efficient ($H_{passive} / H_{max} \approx 0.25 $).

As discussed in section #, high-variance open-loop intervention tends to increase correlations between pairs of nodes downstream of the intervention, and decreases correlations when only one node is downstream of the stimulus location. This can produce more distinct, hypothesis-specific patterns of pairwise dependence. `Fig. DISAMBIG, row Bb` shows how open-loop intervention at node A distinguishes hypotheses \(\{C_1,C_2,C_3\}\) (where node A has reachability to nodes B and C) from hypotheses \(\{C_4,C_5\}\) (where node A can only reach node C). This increased distinguishability is reflected in the distribution of correlation patterns in the fingerprint, and the entropy of that distribution \((H_{OL→A} \approx 1.46 \text{bits}, H_{OL→A}/H_{max} = 0.56)\). In expectation, this intervention provides more information about the hypothesis set than passive observation alone.

For some sets of circuit hypotheses, the capability of closed-loop intervention to remove indirect connections uncovers distinct patterns of resulting correlations that would otherwise be equivalent under other interventions. Because \(C_4\) and \(C_5\) have equivalent reachability matrices, their pairwise correlations will be similar even under open-loop intervention. But in `Row Bb`, closed-loop intervention at node A, severs the inputs to this node. Under hypothesis \(C_4\), nodes C and B remain correlated through their direct connection, however, under \(C_5\), severing inputs to A also severs the indirect influence of C on B, which is sufficient to remove the correlation between nodes C and B. The distribution of observed patterns `(Row Dc)`, contains more distinct entries, and leads to a higher

across-hypothesis entropy of $H_{CL→A} \approx 1.79 \text{bits}$, $H_{CL→A}/H_{max} = 0.69$.

This example highlighted a location for intervention where closed-loop control provides a categorical for distinguishing circuit hypotheses above open-loop control (and passive observation). This advantage is notable, in that it represents an improvement in circuit estimation bias which would be unlikely to be mitigated through collecting more data. However, `Fig. DISAMBIG` further highlights the importance of not only intervention *type*, but also intervention *location* in determining successful circuit inference. For a given intervention type, different locations for delivering stimuli result in categorically different hypothesis-narrowing information *(e.g. $H(OL_B) < H(OL_A) < H(OL_C)$, Fig.* `DISAMBIG` *Column D)*. On the other hand, for interventions at nodes B and C, open-loop and closed-loop control result in identical correlation fingerprints for this hypothesis set — closed-loop control at these locations does *not* provide a categorical benefit beyond the information learned through open-loop control. This equivalence between open-loop and closed-loop interventions arises in cases where severing inputs at the target node does not interrupt an indirect connection which otherwise makes circuits in the hypothesis set ambiguous.

To summarize, by understanding the relationship between circuit structure, the effect of interventions, and changes to the observed patterns of correlation, we were able to demonstrate the relative utility of passive observation, open-loop control, and closed-loop control. Open-loop control improves the capacity to distinguish circuits by increasing the diversity of outcomes as correlations increase or decrease. In addition, closed-loop control is capable of providing a categorical improvement in the ability to distinguish between and narrow down a set of competing hypotheses. It results in distinct patterns of observed dependence in additional cases even with equivalent reachability by severing ambiguous indirect connections. These categorical differences in across-circuit entropy are likely to reflect fundamental differences in the best-case conditions for evaluating similar hypotheses, regardless of data volume or algorithms used for circuit inference.

However, the utility of a given intervention does depend strongly on the location of control relative to paths in the hypothesized circuits. Circuits and hypothesis sets where closed-loop is likely to outperform open-loop control would consist of similar circuits, where direct and indirect connections are difficult to distinguish, such as those with recurrent loops. In highly sparse or largely-feedforward circuits, open-loop and closed-loop intervention are likely to result in similar circuit information.

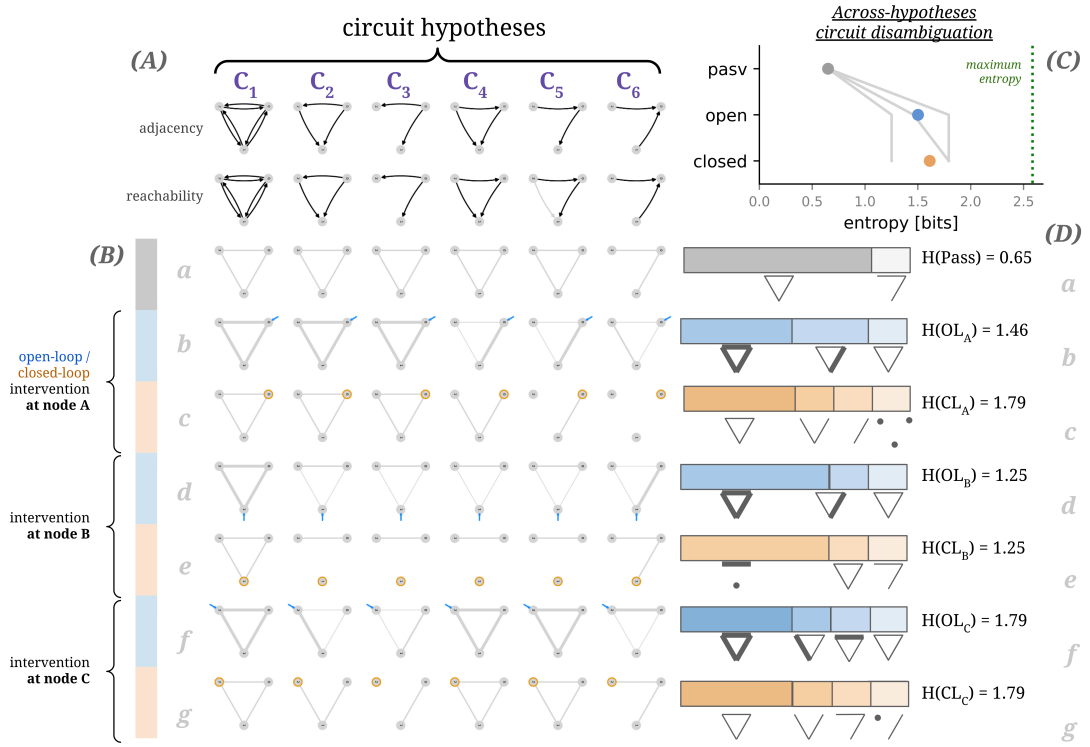


Figure DISAMBIG: Interventions narrow the set of hypotheses consistent with observed correlations

(A) Directed adjacency matrices represent the true and hypothesized causal circuit structure. 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. **(B)** Correlations between pairs of nodes. **a)** Under passive observation, the direction of influence is difficult to ascertain. **(B b-g)** 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[15] edges denote correlations which increase above their baseline value with high variance open-loop input. Thin blue 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. **(C)** Across-circuit entropy for each intervention type and location. Grey lines correspond to a single intervention location. Circle markers represent the mean entropy for a given intervention type across all intervention locations. Green dotted lines represents the maximum achievable entropy for this hypothesis set. **(D)** Distributions of patterns of pairwise correlation across hypotheses, for each intervention location and type. Distributions with more observed patterns, and more uniform probabilities correspond to experiments which reveal more information to narrow the set of candidate hypotheses.

```
text for choosing an intervention ...
```

Stronger intervention shapes correlation, resulting in more data-efficient inference with less bias - *bidirectional var control* (5.1.2)

Impact of intervention location and variance on pairwise correlations (5.1.2.1)

Discussion

```
Restate themes!
- narrowing search space
- where you intervene matters
```

limitations

The examples explored in this work simplify several key features that may have relevant contributions to circuit identification in practical experiments.

[…]

`full observability`

results summary → summary of value closed-loop generally

Closed-loop control has the disadvantages of being more complex to implement and requires specialized real-time hardware and software, however it has been shown to have multifaceted usefulness in clinical and basic science applications. Here we focused on two advantages in particular; First, the capacity for functional lesioning which (reversibly) severs inputs to nodes and second, closed-loop control's capacity to precisely shape variance across nodes. Both of these advantages facilitate opportunities for closed-loop intervention to reveal more circuit structure than passive observation or even open-loop experiments.

summary of guidelines for experimenters

In studying the utility of various intervention for circuit inference we arrived at a few general guidelines which may assist experimental neuroscientists in designing the right intervention for the quesiton at hand. First, more ambiguous hypotheses sets require "stronger" interventions to distinguish. Open-loop intervention may be sufficient to determine directionality of functional relationships, but as larger numbers of similar hypotheses […] closed-loop intervention reduces the hypothesis set more efficiently. Second, we find that dense networks with strong reciprocal connections tend to result in many equivalent circuit hypotheses, but that well-placed closed-loop control can disrupt loops and simplify correlation structure to be more identifiable.[16] Recurrent loops are a common feature of neural circuit, and represent key opportunities for successful closed-loop intervention. The same is true for circuits with strong indirect correlations

`hidden confounds`

“funnel out”, future work → broad impact

`sequential experimental design`

see limitations_future_work.md

Methods

Modeling network structure and dynamics (4.1) — Simulation Methods

Modeling network structure and dynamics

We sought to understand both general principles (abstracted across particulars of network implementation) as well as some practical considerations introduced by dealing with spikes and synapses.

Stochastic network dynamics

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[17].

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

Time-resolvable interactions

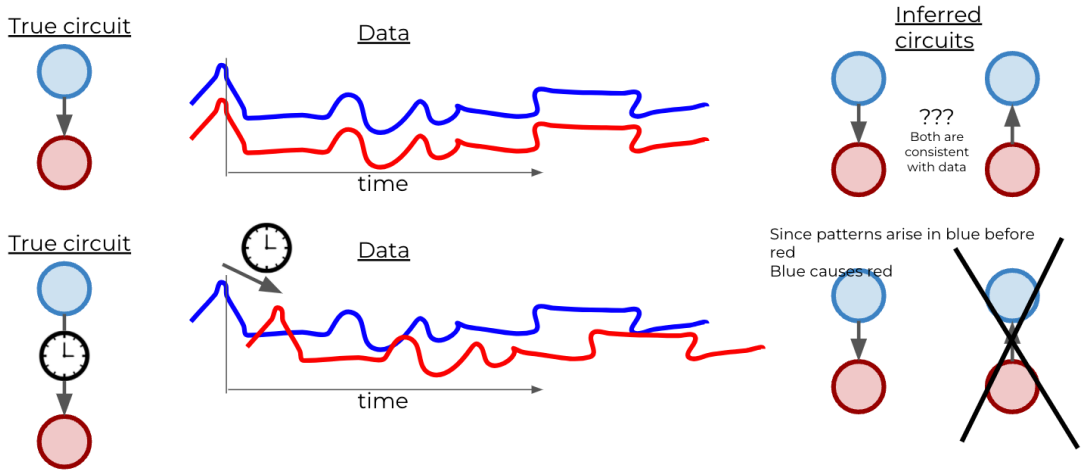
Additionally we study two domains of interactions between populations; contemporaneous and delay-resolvable connections. These domains represent the relative timescales of measurement versus timescale of synaptic delay. [18]

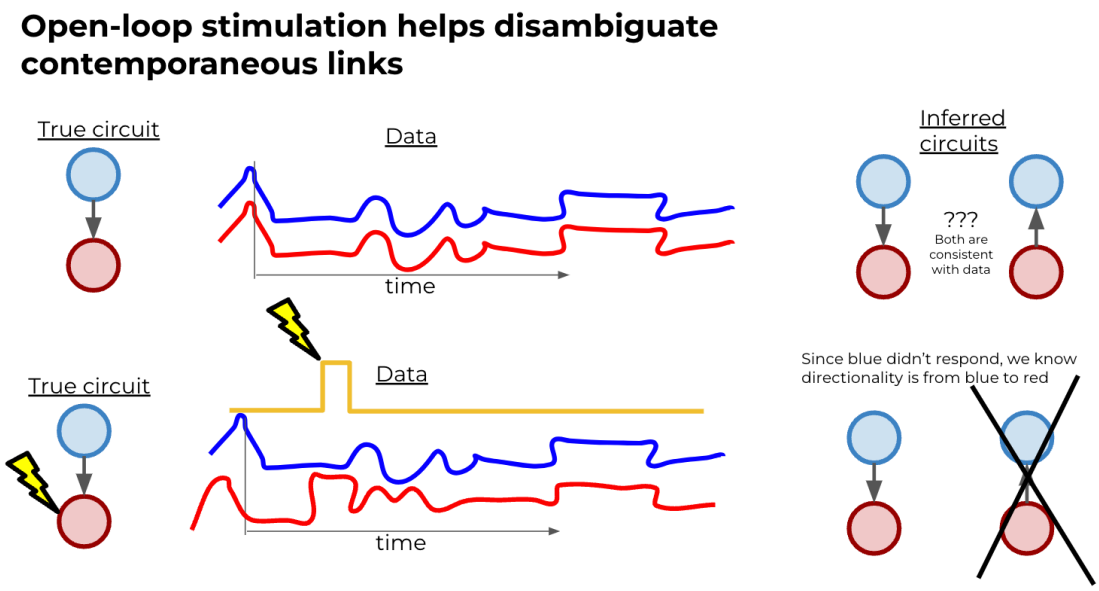
In the delay-resolvable domain, directionality of connections may be inferred even under passive observations by looking at temporal precedence - whether the past of one signal is more strongly correlated with future lags of another signal *(i.e. cross-correlation)*. In the contemporaneous domain, network influences act within the time of a single sample[19] 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.

↪concept figures

But **delayed connections** & observing temporal precedence can identify directionality from **passive observations alone**





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.

see _network_parameters_table.md for list of relevant parameters

Stochastic network dynamics (4.1.1)

Delayed interactions (4.1.2)

Code implementation (4.1.3)

Implementing interventions (4.2)

Implementing interventions

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 $\sigma_m \sqrt{\tau_m}$.[20]

`for the case of leaky integrate and fire neurons:` $$ \frac{dV}{dt} = \frac{V_0 + I - V}{\tau_m} + \sigma_m \sqrt{\tau_m} \xi(t) $$

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 a Gaussian distribution with mean and variance $\mu_{intv.}$ and $\sigma^2_{intv.}$[21]

$$ I_{open-loop} \sim \mathcal{N}(\mu_{intv.},\,\sigma^{2}_{intv.})\\ $$ 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 T. $$ \begin{aligned} T &\sim \mathcal{N}(\mu_{intv.},\,\sigma^{2}_{intv.}) \\ I_{closed-loop} &= f(X, T) \\ X_k | CL_{k} &\approx T \end{aligned} $$ 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[22]. 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

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

↪out of scope: full-loop discrete-time simulation
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.

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

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.

Predicting correlation structure (3.1) — Theory / Prediction

Representations & reachability (2.3?)

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[23].

When the system is linear we can use matrix notation to describe the impact of each node on the others:[24] $$x_{t+1} = W x_t + e_t,$$ where $x_t \in \mathbb{R}^p$ denotes the state of each of the p nodes at time t, and $e_t \in \mathbb{R}^p$ denotes the instantiation of each node's (independent and identically-distributed) private noise variance at time t.

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^{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.

```
see also:
@ import "/section_content/background_id_demo.md"
```

Predicting correlation structure (3.1)

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

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 \(\left[\sum_{k=0}^{\infty} (W^T)^k s \right]_j\).

Our goal is to connect private variances and connection strengths to observed pairwise correlations in the circuit. Defining \(X \in \mathbb{R}^{p

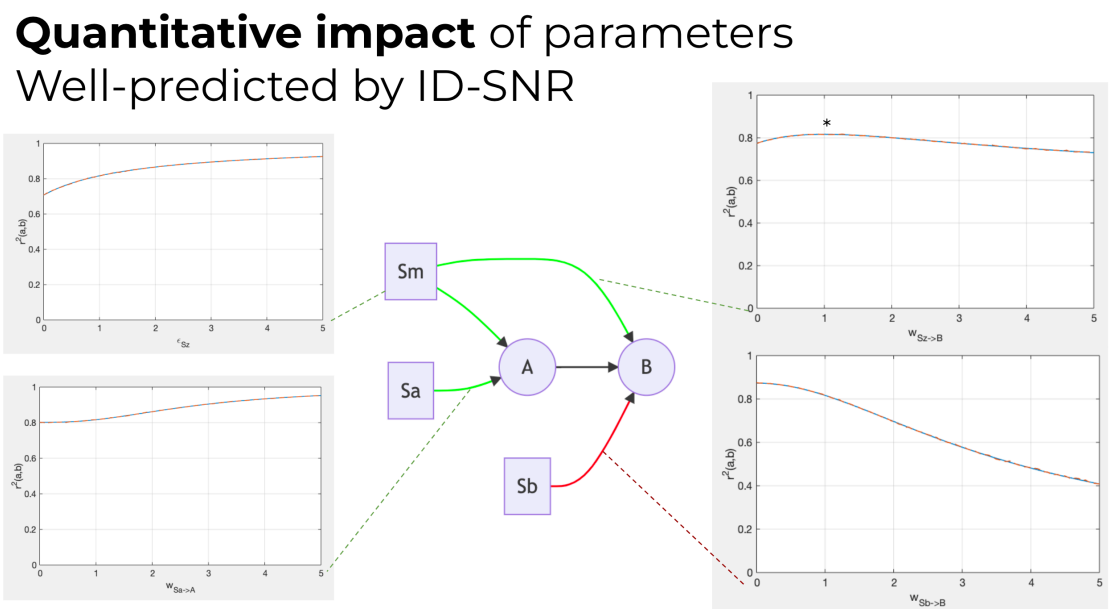
\times n}\) as the matrix of \(n\) observations of each node, we have[25] \[\begin{aligned} \Sigma &= \mathrm{cov}(X) = \mathbb{E}\left[X X^T\right] \\ &= (I-W^T)^{-1} \mathrm{diag}(s) (I-W^T)^{-T} \\ &= \widetilde{W} \mathrm{diag}(s) \widetilde{W}^T, \end{aligned} \] where \(\widetilde{W} = \sum_{k=0}^{\infty} (W)^k\) denotes the *weighted reachability matrix*, whose \((i,j)^\mathrm{th}\) entry indicates the total influence of node \(i\) on node \(j\) through both direct and indirect connections.[26] 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 \[\begin{aligned} r^2(i,j) &= \frac{\Sigma_{ij}}{\Sigma_{ii} \Sigma_{jj}} \\ &= \frac{\left(\sum_{k=1}^p \widetilde{W}_{ik} \widetilde{W}_{jk} s_k \right)^2}{\left(\sum_{k=1}^p \widetilde{W}_{ik}^2 s_k\right)\left(\sum_{k=1}^p \widetilde{W}_{jk}^2 s_k\right)}. \end{aligned} \]

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.[27] Because \(c\)'s inputs have been severed, this private noise will become exactly node \(c\)'s output variance.

The impact of intervention on correlations can be understood from the intervention's location relative to causal circuit connections. One useful distillation of this concept is to understand the sign of \(\frac{dr^2_{ij}}{dS_k}\), 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.



`if:` \[\text{Reach}(S_k→i) \neq 0 \\ \text{Reach}(i→j) \neq 0\] `then:` \[\frac{dr^2}{dS_k} > 0\]

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

`if:` \[\text{Reach}(S_k → j) = 0 \\ \text{Reach}(S_k → j) \neq 0 \] `then:` \[\frac{dr^2}{dS_k} < 0\]

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

`if:` \[\text{Reach}(S_k → i) \neq 0 \\ \text{Reach}(S_k → j) \neq 0 \\ \text{Reach}(i → j) = 0 \] `then:` \[\frac{dr^2}{dS_k} > 0\]

Notably, the impact of an intervention which is a “common cause” for both nodes depends on the relative weighted reachability between the source and each of the nodes. Correlations induced by a common cause are maximized when the input to each node is equal, that is $\widetilde{W}_{S_k→i} \approx \widetilde{W}_{S_k→j}$ (upper right * in fig. variance). If i→j are connected $\widetilde{W}_{S_k→i} \gg \widetilde{W}_{S_k→j}$ results in an variance-correlation relationship similar to the “upstream source” case (increasing source variance increases correlation $\frac{dr^2}{dS_k} > 0$), while $\widetilde{W}_{S_k→i} \ll \widetilde{W}_{S_k→j}$ results in a relationship similar to the “downstream source” case ($\frac{dr^2}{dS_k} < 0$)[28]

Impact of interventions - theory, pred (3.1?, 5.1?)

---

[29]$$\mathbb{V}_{i}(C|S=\text{open},\sigma^2_S) \geq \mathbb{V}_{i}(C)$$
More specifically, if the open-loop stimulus is statistically independent from the intrinsic variability[30] $$\mathbb{V}_{i}(C|S=\text{open},\sigma^2_S) = \mathbb{V}_{i}(C) + \sigma^2_S$$ Applying closed-loop to a linear Gaussian circuit:

$$\begin{aligned} \mathbb{V}_{i}(C|S=\text{closed},\sigma^2_S) &= \sigma^2_S \\ \mathbb{V}_{i}(C|S=\text{closed},\sigma^2_S) &\perp \mathbb{V}_{i}(C) \end{aligned}$$

↪ Firing rates couple mean and variance
In neural circuits, we’re often interested in firing rates, which are non-negative. This particular output nonlinearity means that the linear Gaussian assumptions do not hold, especially in the presence of strong inhibitory inputs. In this setting, firing rate variability is coupled to its mean rate; Under a homoeneous-rate Poisson assumption, mean firing rate and firing rate variability would be proportional. With inhibitory inputs, open-loop stimulus can drive firing rates low enough to reduce their variability. Here, feedback control still provides an advantage in being able to control the mean and

variance of firing rates independently[31]

$$ \begin{aligned} \mu^{out}_i &= f(\mu^{in}_i, \mathbb{V}^{in}_i)\\ \mathbb{V}^{out}_{i}(C) &= f(\mu^{out}_i, \mathbb{V}^{in}_i) \end{aligned} $$

↪ Notes on imperfect control `Ideal control` $$ \mathbb{V}_{i}(C|S=\text{closed},\sigma^2_S) = \sigma^2_S $$ `Imperfect control` - intuitively feedback control is counteracting / subtracting disturbance due to unobserved sources, including intrinsic variability. We could summarize the effectiveness of closed-loop disturbance rejection with a scalar $0\leq\gamma\leq1$ $$ \mathbb{V}_{i}(C|S=\text{closed},\sigma^2_S) = \mathbb{V}_{i}(C) - \gamma\mathbb{V}_{i}(C) + \sigma^2_S \\ \mathbb{V}_{i}(C|S=\text{closed},\sigma^2_S) = (1-\gamma)\mathbb{V}_{i}(C) + \sigma^2_S $$

```
see also:
@ import "/section_content/methods_predicting_correlation.md"
@ import "/section_content/results_impact_of_intervention.md"`
```

Extracting circuit estimates (4.3)

Extracting circuit estimates

!!!! - 10% done

> *refer to methods overview figure*

While a broad range of techniques[32] 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[33] 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.”[34] 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.[35]

Time-resolvable interactions *XCORR* (4.1.2)

`@ import "/section_content/methods_simulations.md" time-resolvable domain`

Information-theoretic measures of hypothesis ambiguity (4.4)

Shannon entropy provides a scalar summarizing the diversity of a set of outcomes.. …how uniform a discrete probability function is… …how surprising…(in expectation)

\[H(X) = E[I(X)] = E[\log\frac{1}{p(X)}] = \sum_{i=1}^{N} p(x_i) \log\frac{1}{p(x_i)} \]

`interpretting high and low entropy`

> a highly predictable experimental outcome means an experiment where not much was learned

An intervention associated with a higher entropy across circuits will, on average, provide more information to narrow the set of hypotheses. In fact, one interpretation of entropy is that it describes the (uncertainty associated with the equivalent) number of *equally-likely* outcomes[36] of a probability mass function. In this setting \(N_{equal}\) can be thought of as the number of hypotheses that can be distinguished under a given experiment[37]. \[H(C)

= \log_2 N_{equal} \\ N_{equal} = 2^{H(C)}\] For instance, open-loop intervention at node \(x_0\) in (Fig.DISAMBIG right column) results in an entropy across the hypotheses of \(H(C|S_0) \approx 1.5\)bits or \(N_{equal} \approx 2.8\). Looking at the patterns of correlation, there are \(N=3\) distinct patterns, with the +++ pattern somewhat more likely than the others (+–, 0–).[38] This intuition also helps understand the maximum entropy achievable for a given set of hypotheses: \[H^{max}(C) = log_2 N\] for this example set: \[H^{max}(C) = log_2 6 \approx 2.6\]

Selecting interventions (…)

> Evolution of entropy, as the space of hypotheses is narrowed from experiments and inference.

\[\begin{align} H^{pre}(C):& \text{ uncertainty before intervention (starts at}\, H^{max}(C))\\ H(C|S_i):& \text{ expected information gain from a given intervention}\\ H^{post}(C|S_i) = H^{pre} - H(C|S_i):& \text{ expected remaining uncertainty after intervention} \end{align} \] If \(H(X|S_i)\approx0 \,\forall i\), none of the candidate interventions provide additional information, and the identification process has converged. If \(H^{post} = 0\) the initial hypothesis set has been reduced down to a single circuit hypothesis consistent with the observed data[39]. If \(H^{post} > 0\), some uncertainty remains in the posterior belief over the hypotheses. In this case a Maximum A Posteriori (MAP) estimate could be chosen as: \[\hat{c}_{\text{MAP}} = \underset{c}{\text{argmax}} \,L(\text{Corr} | c)\,\pi(c) \] or the posterior belief can be used as a prior for the next iteration.

References

see pandoc pandoc-citations

Supplement

"Adam-to-Do." 2022.

Ay, Nihat, and Daniel Polani. 2008. "Information Flows in Causal Networks." *Advances in Complex Systems* 11 (01): 17–41. https://doi.org/10.1142/S0219525908001465.

Barnett, Lionel, Adam B. Barrett, and Anil K. Seth. 2009. "Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables." *Physical Review Letters* 103 (23): 238701. https://doi.org/10.1103/PhysRevLett.103.238701.

Bossomaier, Terry, Lionel Barnett, Michael Harré, and Joseph T. Lizier. 2016. "Transfer Entropy." In *An Introduction to Transfer Entropy*, 65–95. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43222-9_4.

Chicharro, Daniel, and Anders Ledberg. 2012. "When Two Become One: The Limits of Causality Analysis of Brain Dynamics." Edited by Thomas Wennekers. *PLoS ONE* 7 (3): e32466. https://doi.org/10.1371/journal.pone.0032466.

Chis, Oana-Teodora, Julio R. Banga, and Eva Balsa-Canto. 2011. "Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods." Edited by Johannes Jaeger. *PLoS ONE* 6 (11): e27755. https://doi.org/10.1371/journal.pone.0027755.

Das, Abhranil, and Ila R. Fiete. 2020. "Systematic Errors in Connectivity Inferred from Activity in Strongly Recurrent Networks." *Nature Neuroscience* 23 (10): 1286–96. https://doi.org/10.1038/s41593-020-0699-2.

Dean, Roger T., and William T. M. Dunsmuir. 2016. "Dangers and Uses of Cross-Correlation in Analyzing Time Series in Perception, Performance, Movement, and Neuroscience: The Importance of Constructing Transfer Function Autoregressive Models." *Behavior Research Methods* 48 (2): 783–802. https://doi.org/10.3758/s13428-015-0611-2.

Eberhardt, Frederick, and Richard Scheines. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74 (5): 981–95. https://doi.org/10.1086/525638.

Garofalo, Matteo, Thierry Nieus, Paolo Massobrio, and Sergio Martinoia. 2009. "Evaluation of the Performance of Information Theory-Based Methods

and Cross-Correlation to Estimate the Functional Connectivity in Cortical Networks." *PLOS ONE* 4 (8): e6482. https://doi.org/10.1371/journal.pone.0006482.

Ghassami, AmirEmad, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. 2018. "Budgeted Experiment Design for Causal Structure Learning." In *Proc. 35th Int. Conf. On Machine Learning*. Stockholm, Sweden.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37 (3): 424. https://doi.org/10.2307/1912791.

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2013. "Quantifying Causal Influences." *The Annals of Statistics* 41 (5). https://doi.org/10.1214/13-AOS1145.

Knox, Charles K. 1981. "Detection of Neuronal Interactions Using Correlation Analysis." *Trends in Neurosciences* 4 (January): 222–25. https://doi.org/10.1016/0166-2236(81)90070-9.

Lizier, J. T., and M. Prokopenko. 2010. "Differentiating Information Transfer and Causal Effect." *The European Physical Journal B* 73 (4): 605–15. https://doi.org/10.1140/epjb/e2010-00034-5.

Maathuis, Marloes H., and Preetam Nandy. 2016. "A Review of Some Recent Advances in Causal Inference." In *Handbook of Big Data*, 387–408.

Melssen, W. J., and W. J. M. Epping. 1987. "Detection and Estimation of Neural Connectivity Based on Crosscorrelation Analysis." *Biological Cybernetics* 57 (6): 403–14. https://doi.org/10.1007/BF00354985.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second. Cambridge University Press.

Penfield, W., and E. Boldrey. 1937. "Somatic Motor and Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation." *Brain: A Journal of Neurology* 60: 389–443. https://doi.org/10.1093/brain/60.4.389.

Penfield, Wilder, and Theodore Rasmussen. 1950. *The Cerebral Cortex of Man; a Clinical Study of Localization of Function.* The Cerebral Cortex of Man; a Clinical Study of Localization of Function. Oxford, England: Macmillan.

Runge, J. 2018. "Causal Network Reconstruction from Time Series: From Theoretical Assumptions to Practical Estimation." *Chaos: An Interdisciplinary*

Journal of Nonlinear Science 28 (7): 075310. https://doi.org/10.1063/1.5025050.
Salinas, Emilio, and Terrence J. Sejnowski. 2001. “Correlated Neuronal Activity and the Flow of Neural Information.” *Nature Reviews Neuroscience* 2 (8): 539–50. https://doi.org/10.1038/35086012.
Schreiber, Thomas. 2000. “Measuring Information Transfer.” *Physical Review Letters* 85 (2): 461–64. https://doi.org/10.1103/PhysRevLett.85.461.
Shorten, David P., Richard E. Spinney, and Joseph T. Lizier. 2021. “Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains or Other Event-Based Data.” *PLOS Computational Biology* 17 (4): e1008054. https://doi.org/10.1371/journal.pcbi.1008054.
Wibral, Michael, Raul Vicente, and Joseph T. Lizier, eds. 2014. *Directed Information Measures in Neuroscience*. Understanding Complex Systems. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54474-3.
Wiener, N. 1956. “The Theory of Prediction.” In *Modern Mathematics for the Engineer*. McGraw-Hill.
Yang, Karren D., Abigail Katoff, and Caroline Uhler. 2018. “Characterizing and Learning Equivalence Classes of Causal DAGs Under Interventions.” In *Proc. 35th Int. Conf. On Machine Learning*. Stockholm, Sweden.

---

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. TODO: need a more accurate summary of types of models we look at.↩
3. TODO: need to assess total scope, cut or diminish reference to time-lagged correlations if it doesn’t make it to final paper↩
4. fill out rest of intervention caption↩
5. 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.↩

6. if citations needed here, could start by looking for a good high-level reference in either (Ghassami et al. 2018) or (Yang, Katoff, and Uhler 2018). (Both of these papers are pretty technical, so likely wouln't be great citations on their own.)↩
7. TODO: make sure this citation is in the right place)↩
8. saying "difficult to distinguish" instead of "indistinguishable" here since the magnitudes of the correlations could also be informative with different assumptions↩
9. see Advancing functional connectivity, fig. 2↩
10. more than just an experiment, this is a "hypothesis search." Is this procedure what we're going to brand as the "CLINC" process?↩
11. verify whether this is reasonable to say↩
12. using binary reachability, we can be more general above predicting the "sign/slope" (when will they increase/decrease) of other measures of bivariate dependence like transfer entropy↩
13. will need to tighten up notation for intervention summarized as a variable, annotating its type (passive, open-, closed-loop) as well as its location. Also have to be careful about overloading \(S_i\) as the impact of private variance and as a particular open-loop intervention↩
14. Omitting several quantitative practicalities in this step. Notably choosing the amplitude / frequency content of an intervention w.r.t. estimated parameters of the circuit↩
15. 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↩
16. this corroborates Ila Fiete's paper on bias as a function of recurrent network strength↩
17. However, depending on overall firing rates and population sizes, this sparse spike-based transmission can be coarse-grained to a firing-rate-based model.↩
18. cases doesnt work with pandoc yet, also want to talk about positive and negative lags here↩
19. the effective \(\Delta_{sample}\) would be broadened in the

presence of jitter in connection delay, measurement noise, or temporal smoothing applied post-hoc, leading↩

20. need to triple check indexing w.r.t. nodes, neurons↩
21. need to resolve differences in implementation between contemporaneous and voltage simulation cases↩
22. NEED dynamic clamp refs - http://www.scholarpedia.org/article/Dynamic_clamp↩
23. 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.↩
24. 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↩
25. 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 $\Sigma = \mathrm{cov}(X) = \mathbb{E}\left[X X^T\right]$ can then be written as $\Sigma = \mathbb{E}\left[(I-W^T)^{-1} E E^T (I-W^T)^{-1} \right] = (I-W^T)^{-1} \mathrm{cov}(E) (I-W^T)^{-T} = (I-W^T)^{-1} \mathrm{diag}(s) (I-W^T)^{-T}$.↩
26. 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.↩
27. TODO: to any target value?↩
28. TODO: verify not 100% sure this is true, the empirical results are really pointing to dr^2/dW<0 rather than dr^2/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.↩

29. need to be clear V means variance↩
30. notably, this is part of the definition of open-loop intervention↩
31. 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.↩
32. *inference techniques mentioned in the intro…*↩
33. what does "prototype" mean here? something like MI and corr are equivalent in the linear Gaussian case, …↩
34. TODO? formalize notation for this↩
35. 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↩
36. i.e. if you took a PMF and counted the number of categories with probability greater than \(p_th\). A distribution with 16 possible outcomes, but only 2bits of uncertainty is *as* uncertain as a uniform distribution with \(2^2\) equally likely outcomes↩
37. connect this section to the idea of the markdov equivalence class, and its size↩
38. since \(H(C)\leq H^{max}(C)\), \(N_{equal} \leq N\)↩
39. what about the scenario where the ground truth circuit is not in the hypotheses set?↩