Article type: Perspective

Title:
Advancing functional connectivity research from association to causation

Authors:
Andrew T. Reid[1], Drew B. Headley[2], Ravi D. Mill[2], Ruben Sanchez-Romero[2], Lucina Q. Uddin[3], Daniele Marinazzo[4], Daniel J. Lurie[5], Pedro A. Valdés-Sosa[6], Stephen José Hanson[7], Bharat B. Biswal[8], Vince Calhoun[9], Russell A. Poldrack[10], Michael W. Cole[2]

Affiliations:
[1]School of Psychology, University of Nottingham, Nottingham, NG7 2RD, United Kingdom
[2]Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ, 07102, USA
[3]Department of Psychology, University of Miami, Coral Gables, FL 33124, USA and Neuroscience Program, University of Miami Miller School of Medicine, Miami, FL 33136, USA
[4]Department of Data Analysis, Ghent University, Ghent, 9000, Belgium
[5]Department of Psychology, University of California, Berkeley, Berkeley, CA, 94720, USA
[6]The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, 2006, Xiyuan Ave. West Hi-Tech Zone, 611731 Chengdu China; Cuban Neuroscience Center, La Habana, Cuba
[7]RUBIC & Department of Psychology, Rutgers University, Newark, NJ, 07102, USA
[8]Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, 07102, USA
[9]Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) [Georgia State University, Georgia Institute of Technology, Emory University], Atlanta, GA, 30303
[10]Department of Psychology, Stanford University, Stanford, CA, USA

Corresponding author:
Michael W. Cole
Center for Molecular and Behavioral Neuroscience
197 University Ave, Suite 212
Newark, NJ 07102
973-353-3249
mwcole@mwcole.net

Abstract

Cognition and behavior emerge from brain network interactions, such that investigating causal interactions should be central to the study of brain function. Approaches that characterize statistical associations among neural time series – functional connectivity (FC) methods – are likely a good starting point for estimating brain network interactions. Yet only a subset of FC methods ("effective connectivity") are explicitly designed to infer causal interactions from statistical associations. Here we incorporate best practices from diverse areas of FC research to illustrate how FC methods can be refined to improve inferences about neural mechanisms, with properties of causal neural interactions as a common ontology to facilitate cumulative progress across FC approaches. We further demonstrate how the most common FC measures (correlation and coherence) reduce the set of likely causal models, facilitating causal inferences despite major limitations. Alternative FC measures are suggested to immediately start improving causal inferences beyond these common FC measures.

Introduction

The basic physical attributes of brain networks have been extensively characterized, yet the functional units (e.g., neurons, brain regions) and the dynamics that produce cognition and behavior remain poorly understood. We suggest that a focus on brain network interactions – often termed functional connectivity (FC) – may be a primary means of understanding brain function across these levels of organization, by identifying interactions at each level and ultimately how those interactions produce cognition.

Yet FC as it is currently defined suffers from a variety of theoretical and practical issues that limit its ability to advance neuroscientific understanding. In this Perspective we identify these issues and propose a framework to begin remedying them (**Table 1**). This framework will appear familiar to experienced FC researchers, as it incorporates insights and best practices from FC research approaches (including "effective connectivity") and beyond. Nevertheless, we expect it to be useful for both novice and expert FC researchers, due to insights gained from integrating across typically-separate areas of FC research. We build on previous discussions on this topic[1–6] with the goal of making this debate more accessible and suggesting a novel way forward. As interactions among neural units are central to neurocognitive function, we anticipate fundamental improvements in FC theory and methodology will have widespread benefits for advancing neuroscience.

Perhaps the most fundamental issue with FC in its current conceptualization is how it is typically defined: as the statistical association between measured brain signals[7,8]. This is problematic because it fails to distinguish target theoretical properties of interest from the methods used to infer those properties. This is akin to defining the moon as the photons that hit one's retina when looking at a particular location in the sky (a common method for detecting the moon), rather than as a physical object with a variety of properties consistent with the laws of physics (theoretical properties of interest). In other words, it confuses the map with the territory – a classic logical fallacy[9] that impedes scientific progress. As these are issues with fundamental

scientific inferences the framework (Table 1) is applicable to a variety of scientific problems, though we emphasize its application to FC research here.

Generalizing insights from existing FC approaches[1,2], we propose that the ultimate phenomenon of theoretical interest in all FC research is understanding the causal interaction among neural entities. This clearly runs counter to the typical definition of FC as the non-causal "statistical association" between measured brain signals. Nevertheless, it is in line with the kinds of inferences that should be sought in FC research, given that a physical means of interaction between neural entities is implied by the term "connectivity" (as in structural/axonal connectivity). Further, FC researchers already work within a causal inference framework, whether they realize it or not. For instance, FC is often used to identify sets of correlated brain regions that are then commonly treated as real causal entities (physical systems) known as large-scale brain networks or brain systems[10–13]. Additionally, when it was discovered that in-scanner motion was strongly associated with fMRI-based FC estimates, this was generally treated as a causal problem, with motion as an alternate causal path confounding proper FC inference[14–16].

This tendency to already interpret FC measures in a causal framework suggests it would be natural to elevate causal reasoning from implicit to explicit in FC research. The kinds of causal inferences that can be made using FC methods is often limited, however, because simple statistical measures such as coherence and Pearson correlation allow ambiguous causal paths among measured neural signals. Despite their limitations, we illustrate below how even simple FC measures can be informative (albeit often weakly) with regard to causal inferences. Notably, FC measures labeled "effective connectivity" can often enable more precise causal inferences[2,6], though they are not without their own limitations (see below). Despite these limitations (and in contrast to some others[17]) we view any narrowing of the space of likely causal graphs as progress toward the ultimate goal of strong causal inferences in FC research.

Our proposal to shift the focus of FC from association to causal interaction derives from many considerations, though in large part from the increased confidence in defining causality and making causal inferences that has coincided with what has been called the "Causal Revolution" occurring over the past 25 years[18–22]. Central to this increased confidence have been demonstrations of new methods to make valid causal inferences from observational data, expanding causal inference beyond the limited purview of randomized controlled experiments[19,23].

Especially transformative to progress in defining causality is the realization (building on centuries of work in philosophy[24]) that counterfactuals (e.g., experimental control) can be used to conclusively define causality[19,25]. As a particularly clear example of counterfactual causality in science, the concept of a controlled experiment implicitly invokes this definition: Comparing a treatment condition (with the cause present) with a control condition (with the cause absent) informs us what would have happened in the treatment condition had the treatment not been applied. Accordingly, we define the "cause" of an observed neural event (the "effect") as a preceding neural event whose occurrence is necessary to observe the effect. Causality can thus be demonstrated by observing a system under two conditions, differing only in the presence or absence of the causing event. Note that this focus on neural interactions is a more circumscribed definition of causality than the cognition-focused causation typically used in lesion and stimulation studies in neuroscience (i.e., that activity in a neural entity is necessary for a cognitive function). Even with this more modest goal for

causal inference, making such inferences is complicated by methodological limitations. Making progress despite these limitations is a major focus of this Perspective.

One solution to the problems faced by FC research might be to abandon the term "functional connectivity" altogether – an idea espoused by at least one of the authors (PAVS). Going beyond issues with terminology, our primary goal in developing the current framework is to create a unifying conceptualization of FC, accommodating both methodological and target theoretical properties using the logic of causal inference. One prominent divide among FC methods is the supposed distinction between "effective connectivity" and other FC approaches[7]. Unfortunately, there is some confusion in the field over how to define the concept of effective connectivity (although attempts have been made to address this[1,2]), with emphasis sometimes placed on the target theoretical property of whether a connection is direct vs. indirect (e.g., via a third brain region)[7,26,27] or whether a connection is directed vs. undirected/bidirectional[2,7,28]. We seek to remedy this situation by placing all such methods under the umbrella of FC, with a systematic taxonomy that clarifies what theoretical properties each FC method targets, and hence what aspects of brain network interactions each is capable of characterizing.

Another major goal of developing an FC-focused framework is to bring best practices for grounding FC findings in physical mechanisms into focus. Bridging the gap between FC observations and physical mechanisms is clearly easier in some cases (e.g., invasive animal models) than others (e.g., noninvasive methods such as fMRI and EEG). This is in part due to the indirect nature of methods like fMRI and EEG, which introduces ambiguities into interpretation (**Figure 1**). We frame this problem as a matter of mapping observations to a hypothesis search space[29,30] consisting of different possible causal interactions among neural entities, with each observation constraining the likelihood of each hypothesis. With proper validation of FC methods, we suggest it is possible to produce minimally ambiguous interpretations, especially when multiple FC methods are combined to create a convergent interpretation. The framework builds on recent simulation-based and empirical validations of FC measures[31–33] to suggest a way forward for FC method validation, with the goal of making accurate inferences about brain function. We expect that constraining the hypothesis space by seeking convergence across validated methodologies and replications will bring us towards a mechanistic understanding of brain network functions[34].

In the following sections we begin with a summary of the proposed framework. Remaining issues with FC interpretation are then detailed, along with a general strategy for validating mechanistic interpretations of FC methods to help overcome these issues. Suggestions for how to apply these principles to commonly used FC methods (fMRI, EEG, and intracranial recordings) are also provided as Supplementary Material. Together, the proposed framework integrates best practices from across FC research to provide a way toward achieving more valid inferences of the FC properties that are of theoretical interest to the neuroscience community.

Summary of the proposed framework

We propose a framework that incorporates best practices and insights from diverse areas of FC research, targeting three key gaps: 1) the need for an account of FC as both a theoretical and methodological construct; 2) the need to reconcile functional (and effective) connectivity approaches within a single theoretical ontology grounded in biological mechanisms; and 3) a systematic means of validating theoretically-meaningful interpretations of results obtained using FC methods.

The FC framework is a conceptual structure within which a taxonomy of FC methods and relevant inferences can be detailed (see **Box 1** for definitions of key terms). The taxonomy consists of a series of mappings, each between a methodological procedure and the inferences that can be based on it (**Figure 2A**). These inferences are built from three classes of property: 1) *Target theoretical properties*, representing the potential scientific purposes of a method and what inferences of theoretical importance it can support; 2) *Methodological properties*, representing limitations and enhancements imposed by the method that are not of direct theoretical interest for understanding the brain; and 3) *Confounding properties*, alternate (often non-neural) causes of observed effects, which must be addressed to make valid inferences.

As a brief illustration of this framework, consider a statistically significant Pearson correlation between two brain regions based on resting-state fMRI data. A *target theoretical property* could be whether the two regions causally interact at rest. The *target theoretical properties* that can be validly targeted by a given method are limited by *methodological properties* as well as *confounding properties*. For instance, *methodological properties* indicate several ambiguities when using Pearson correlation with fMRI data, which undermine support for the inference. Specifically, it is ambiguous whether the potential interaction, mediated via action potentials over axons, is direct or involves other regions. The direction of the interaction is also ambiguous, as are other properties (e.g., its temporal frequency). Finally, the target inference can be made only when *confounding properties* inconsistent with the target theoretical properties have been controlled for (e.g., correction for motion artifacts). Given these properties, a statistically significant FC result would support the following inference (which is somewhat weaker than the target inference): The two regions interact (directly or indirectly, with ambiguous directionality), and/or share mutual interactions with other regions, during resting state. See Supplemental Material for additional details. Despite the weakness of this inference, it informs our understanding of these two regions by revealing that some causal models are more likely than others (see **Figure 2**). It also points toward the need to use and/or develop better methods for strengthening the intended causal inference.

We next describe how this framework addresses the three core problems with FC research mentioned above.

Problem 1: The need for an account of FC as both a theoretical and methodological construct

Every methodological approach should have one or more theoretical targets if it seeks to be explanatory rather than just descriptive. Unfortunately, the FC literature has often failed to identify the target theoretical construct, or has inferred target theoretical constructs beyond those permitted by methodological properties of a given FC approach. This has led to a tendency to interpret the results of a specific approach in terms of biological mechanisms, when that approach does not warrant this level of interpretation. We address this by framing FC in terms of mapping empirical results to target theoretical properties via methodological properties. This allows us to (1) reduce the temptation to overinterpret findings, and (2) identify the limitations of a given methodology with respect to its ability to support inferences targeting biological mechanisms.

Two uses of FC that have recently become popular – dynamic FC and FC-behavior associations – are particularly illustrative of this problem. In the case of dynamic FC, time-varying changes in FC are often characterized[35,36] without considering the mechanisms driving the FC measures (and their changes) to begin with. Similarly, FC-behavior associations have revealed potential insights into the neural basis of cognition[37–39], yet incomplete understanding of the mechanistic basis of FC also limits the utility of this approach. It will therefore be important to advance causal understanding of FC measures so results involving dynamic FC and FC-behavior associations (along with other uses of FC methodology) provide mechanistic insight into brain function.

It is worth noting that "effective connectivity" approaches, such as dynamic causal modelling (DCM), Granger causality, or Bayesian search methods, are typically clearer about their target theoretical inferences. For instance, DCM restricts FC-related inferences to a particular structural graph and a specific form of directed FC among a set of nodes[40]. Adding assumptions to their modeling of FC helps link observations to these target theoretical properties. However, these assumptions are often by necessity unrealistic or overly simplistic – for instance, only incorporating a small subset of brain regions (and failing to account for extraneous influences), evaluating only unrealistically sparse networks, or modelling properties such as connection weight as a single global parameter[41–43]. The current framework seeks to expand the sort of reasoning underlying effective connectivity approaches to the whole of FC-related research, while acknowledging an inherent trade-off between the competing imperatives of accounting for complexity and potential confounds on the one hand, and modelling the nervous system in its entirety on the other.

Problem 2: The need to integrate functional (and effective) connectivity approaches into a single framework

Sixteen years ago the classic paper "The elusive concept of brain connectivity"[8] made a strong case that connectivity research was not a cumulative scientific enterprise. According to that paper, "Until it is understood what each definition means in terms of an underlying neural substrate, comparisons of functional and/or effective connectivity across studies may appear inconsistent and should be performed with great caution."[8]. The current framework thus seeks to develop a "common currency" for comparison of results across different FC measures. This will allow corroboration of theoretically important results across FC approaches, which may together constrain neurocognitive theories (see **Figure 2A**).

We agree with Horwitz[8] that it is essential to understand the mapping between each FC measure and its neural substrates. As measurements entail different levels of ambiguity, it is important to be explicit about the limitations and assumptions a particular method requires when making such a mapping. Accordingly, our proposed separation of FC properties into three distinct classes (target theoretical, methodological, and confounding) allows inferences about causal interactions to be made at various levels of ambiguity and uncertainty.

We take the position that all FC measures are useful provided they reduce the hypothesis space – the vast set of network configurations that are possible among the theoretical target properties. Thus, rather than seeing ambiguities about target theoretical properties as evidence that a given FC measure is useless or flawed, we focus on what information it provides that helps us constrain neurocognitive theory. For instance, a consistent non-zero Pearson correlation between two regions' fMRI time series increases the probability of a causal interaction existing between those regions, ambiguities about the direction and directness of the relationship notwithstanding (see **Figure 2B,C**). Notably, this strategy is analogous to another that was successfully employed when available evidence was typically ambiguous, in developing and validating the theory of evolution by natural selection[44].

The proposed framework seeks to enumerate a common set of target theoretical FC properties, based on key properties of neural systems. These target theoretical properties are based on the standard model of neural interaction as described by Hodgkin and Huxley[45] and elsewhere[46,47]. We assume that FC measures seek to infer some aspect of causal interaction among neural entities, mediated by action potentials via synaptic transmission. We therefore emphasize target theoretical properties that refer to aggregate action potentials and postsynaptic potentials, as well as the various means to alter the relationship between them (e.g., synaptic strengths, timing). Notably, several researchers have begun combining FC measures to constrain causal graphs of brain region interactions[2,48,49], demonstrating that converging multi-method FC evidence can be used to constrain the hypothesis space (see **Box 2**). A similar approach has also been developed to improve inferences about changes in Pearson correlation-based FC via combination with a simple covariance-based FC measure[4].

Problem 3: The need to validate FC methods to improve mapping of FC results to properties of theoretical interest

Improved validation of FC methods would substantially increase our ability to make strong FC-related theoretical inferences. This reflects the core of the framework: clear mappings between FC method-driven observations and target theoretical inferences. Simulation-based and empirical validations serve to establish these mappings, which can then be generalized to make inferences in new, theoretically-informative scenarios.

We expect the proposed framework to advance efforts to validate FC methods in several ways. First, it clarifies what needs to be validated by explicitly stating the target theoretical properties that should be detected by a given FC method. Second, the framework makes it clear that confounding variables need to be accounted for before a given method can be considered "validated" – ready for use to make target inferences with empirical data. Finally, we flesh out the framework's use of mappings

between observation and theoretical targets to develop strategies for FC method validation.

Details of the FC mechanism framework

Step 1: Identifying target theoretical properties

A mechanism refers to a causal chain of events, and thus for FC our target theoretical properties are, minimally, *causal interactions* between neural entities. Ideally, a causal interaction should be described as having directionality, directness, and weight. *Directionality* refers to the direction of information/activity flow; given neural entities A and B, it specifies whether activity passes from A to B, B to A, or in both directions. *Directness* refers to the number of relays required for activity to pass between A and B; in other words, whether it is a direct (monosynaptic) connection or an indirect (polysynaptic) one. *Weight* refers to the strength of the connection; in other words, how much the signal in A influences the signal in B, as well as whether it is excitatory or inhibitory. In practice, the majority of methods currently used to observe neural activity in humans lack sufficient temporal and/or spatial resolution or coverage to support a full causal description. Yet, they remain useful for supporting weaker causal inferences that might be ambiguous with respect to directionality, directness, or weight.

For FC approaches, we can define "neural entity" as a spatially contiguous region of neural tissue generating a signal (**Box 1**). This encompasses a range of possibilities: small anatomical entities, such as neurons or microcolumns, or larger parcels of neural tissue, whose boundaries are determined cytoarchitectonically or otherwise[10,50–52]. Often, theoretical sources such as current dipoles[53], units of a reference grid such as voxels, or the locations of EEG or intracranial electrodes are also treated as neural entities. In order to support inferences about biological mechanisms, however, each neural entity should describe how its time series integrates action potentials and/or post-synaptic potentials over time and space (**Figure 1**).

Step 2: Identifying methodological properties

It is crucial to be explicit about the methodology employed to obtain the evidence used to support target inferences. As outlined in **Table 1**, this includes several important properties inherent to any observational approach. The *temporal and spatial resolution* of the sampling method constrain how interactions are inferred and how neural entities are defined, respectively. In order to assess temporal precedence, for instance, it is critical to have a sufficient sampling rate to determine the order in which neural entities are activated. Similarly, if the sampling rate is low compared to the connection latency, it can become difficult to determine the directness of observed interactions. Spatial sampling is more critical for the definition of neural entities. This refers to both spatial coverage and spatial resolution; if sensors are too sparsely or too focally distributed there is a risk of failing to capture the complete set of neural entities (and potentially fall victim to confounding; see **Box 2**).

Because the goal of FC approaches is to elucidate biological mechanisms, it is essential to specify how observations map onto their biological causes. This mapping can first be done abstractly, by defining the *observational pathway* through which neuronal activity (action potentials or post-synaptic potentials) maps via sequential levels to the sensors sampling their (typically aggregate) activity (see **Figure 1** for an overview). The observational pathway can then inform the *observation equation*[1], which formally specifies how neural entity states generate the observed signal for a given modality. These equations can range from simple (e.g., spatiotemporal averaging of local field potentials) to highly detailed (e.g., layer-resolved biophysical neural population models)[54]. They depend largely on the nature of the recording apparatus, and represent an integration of existing theoretical knowledge and methodological assumptions about the signal generating process. The physical processes involved in the blood-oxygenation-level-dependent (BOLD) signal[55,56], for instance, are clearly distinct from those generating an extracellular local field potential recorded from an implanted electrode[57]. The observation equation should specify these details, ideally indicating aspects of the observational pathway that remain unknown or are ambiguous.

Finally, we focus on the importance of *assumptions* for describing and interpreting FC approaches and their results. We propose as a key aspect of this framework the enumeration of all critical assumptions required for an FC approach. Specifically, assumptions should be made explicit if they are judged to: (1) be essential for interpreting a methodological result or inference; and/or (2) be uncertain or have the potential to be contentious. Assumptions provide clear focal points for critical discussion, and can be associated either with the analytical methodology, or with the observation equation. For example, in fMRI-based FC studies it is typically assumed that the hemodynamic response function (describing neurovascular coupling) is homogeneous across the brain, or across individuals. The validity of these assumptions is a matter of ongoing debate in the field, however[58–63], and it is therefore important to include them when describing an fMRI-based FC measure requiring them[64]. In general, distinguishing assumptions from the technical details of a given approach can greatly facilitate dialogue addressing that approach and its findings, and failing to do so risks obscuring that dialogue.

Step 3: Identifying confounding properties

We also want to identify all uncontrolled factors that may confound our causal inferences. Formally, a “confounder” refers to any variable that influences two or more variables of interest such that spurious associations arise[65]. For FC analysis, such confounders fall into one of several categories. First, confounders can be non-neural factors that introduce correlated noise simultaneously into multiple neural entities. Examples include physiological artifacts, head motion, and environmental noise. Second, violations of methodological assumptions can also result in confounds. Shared variance between neural entities, for instance, can arise from spatial smoothness induced by image reconstruction, or as a result of source signal mixing in EEG/MEG, which cannot be completely removed through current source localization approaches[66]. Finally, confounds can arise from observed or unobserved neural entities influencing two or more other neural entities (see Box 2).

Having identified potential confounding properties, it is equally important to specify how they will be addressed. Ideally, this would be done by obviating potential confounds prior to data collection. Head motion might be minimized, for example, by means of a head restraining apparatus. Confounding variables can alternatively be accounted for by measuring them directly, and removing their portion of the variance from the neural time series. Physical factors such as head motion or physiological artifacts are commonly addressed in this manner. In cases where confounding factors cannot be directly measured, they can also be isolated via signal decomposition approaches such as independent component analysis, for which artifactual components can be identified and their variance removed[67–71].

If a confound is not addressed via methodological properties, this should be reflected as an ambiguity in the target theoretical properties. This is critical, since the theoretical inference drawn from a given observation must both be supported by its methodological properties, and properly address its confounding properties. This implies that, in the absence of effective control of confounds, we should modify our inferences to explicitly state the possibility that observed effects are due to confounding.

Looking forward: The central role of FC method validation

Valid mappings from FC observations to theoretical properties of interest are critical for gaining mechanistic insights. In this section we provide a systematic approach to validate FC methods (**Figure 3**). Validating mechanistic interpretations of an FC measure involves: (1) identification of a series of ground-truth theoretical and confounding conditions, using either simulation or empirical experiments; and (2) tests of the FC measure for sensitivity to the ground-truth conditions. This can be considered as a series of "forward mappings", from theoretical/confounding properties to FC observations. For example, if a manipulation of the theoretical properties in a particular network configuration can be detected by an FC method, it can be said to be sensitive to that manipulation. Identification of many such forward mappings allows us to quantify our confidence that the method can capture a given property; in other words, it allows us to infer (3) the selectivity of the FC measure for those properties (a "backward mapping"). This is the mapping of interest for future studies: from FC observation to target theoretical properties.

It is critical that the set of sensitivity tests include both plausible confounding properties and target theoretical properties of interest. Without sensitivity testing of confounding properties – along with strategies to minimize or eliminate confounds if sensitivity to them is established – there can be only minimal confidence in the validity of mechanistic interpretations of FC observations. In turn, without sensitivity testing of theoretical properties of interest there can be no valid basis for inferring causal mechanisms from FC observations.

As a simple example of utilizing this validation framework, consider testing of Pearson correlation in a neural mass simulation of fMRI data. Spiking in neural entity A directly and bidirectionally connected to neural entity B causes a non-zero Pearson correlation in the simulated fMRI signal (sensitivity test 1). However, spiking in neural population B also causes non-zero Pearson correlation (sensitivity test 2). From these two tests, we infer the selectivity of the mapping from FC observation to target

theoretical properties: that observing an fMRI Pearson correlation only allows us to infer that an interaction likely occurred between A and B with ambiguous directionality. In practice, more sensitivity tests should be included to test for confounding properties and interactions with additional neural entities.

There are a wide variety of strategies that can be used for sensitivity tests during FC method validation, each with strengths and weaknesses. The basic validation strategies we focus on here are: (1) detailed simulations; (2) abstract simulations; and (3) empirical validations.

Relative to empirical studies, detailed simulations have the advantage that a large number of sensitivity tests can be conducted across the space of possible ground-truth conditions. On the other hand, these simulations typically require many more assumptions. This reflects the complexity of the nervous system; both because our knowledge is imperfect and for computational tractability we must make approximations. This is true even of detailed neuron-level simulation studies[46,47,72,73], for which numerous assumptions are typically necessary to fill gaps in current knowledge and produce plausible neural interactions. One way to overcome this limitation is to vary model parameters over a range of plausible values (e.g., action potential conduction delays of 10-100 ms in 1 ms increments) to ensure the FC method remains valid over this range. Another approach is to focus a model's detail on properties that are most relevant to the method being validated. For example, modeling individual ion channels may be useful for validating a calcium imaging-based approach, but not for an EEG-based approach.

An alternative validation strategy is abstract simulations, for which parameters are reduced to abstract or simplified equivalents[31,74–77]. For instance, rather than modeling every neuron in an entity, one can simulate a "neural mass" that generates averaged neuronal activity. This can have multiple advantages. First, it is easier to intuitively understand abstract models than detailed ones. Second, abstract simulations are much more computationally efficient, which is especially important for large-scale simulations and sensitivity tests across a wide range of ground-truth conditions. Third, an abstract simulation can be equivalent to generalizing over many parameters in a detailed simulation, increasing confidence in the generalizability of the validation results. Despite these advantages, a substantial limitation of abstract simulations is the possibility that an omitted detail would change the outcome of the simulation, resulting in inaccurate sensitivity tests. This could be the case, for instance, if an abstract model were to assume a single large conduction delay between neural units when a given FC method is actually not sensitive to small but realistic conduction delays.

The best strategy to minimize assumptions is to use empirical validation. The absolute ideal would be to know the ground truth in the system of interest (e.g., the human brain) for the context about which you want to make FC inferences. In practice, however, empirical validation involves the limited set of scenarios in which we can establish (or strongly expect) the ground truth and test the sensitivity of an FC method to that ground truth[32,78,79]. The validations in such limited scenarios are then expected to generalize to other scenarios of interest in which ground truth is unknown. As an example, a recent study used the established "memory reactivation" effect in which the portion of sensory cortex representing a sensory experience is reactivated along with that memory[80], to establish a ground-truth reversal of directed FC between visual and auditory cortices in humans[32]. This resulted in evidence that a variety of fMRI and MEG

FC measures are sensitive to the direction of interaction among cortical regions[32]. This validation involved only minimal assumptions relative to simulation-based validation, yet unlike simulations it was limited to only a pair of experimental conditions (involving two brain regions). Other empirical validations of FC measures have involved animal models[77,81–84], allowing for more extensive ground-truth manipulations, but limited by the untested assumption that findings generalize to the human brain.

There have been multiple successful empirical tests for confounding properties. For instance, several strategies have been proposed to establish the empirical ground truth of head movement as a confound for Pearson correlation-based FC with fMRI[14–16]. One such strategy, the comparison of high- versus low-motion subjects, has revealed extensive brain-wide differences in FC estimates[85]. Various strategies have likewise been proposed to address this motion confound. While linear regression was unable to fully address motion effects[85], other approaches such as removing high-movement time points from time series have been effective in making FC estimates more similar between high- and low-motion subjects[86]. Simulation-based validation can also be applied to establish confounding properties and strategies to correct them[4,87], although in such cases it is important to also establish relevance by demonstrating the existence of the confound in empirical data.

Ultimately, it is clear that FC method validation requires convergent evidence across several of these validation strategies. Simulations can provide a broad search over many sensitivity tests to help determine what theoretical FC properties a given FC measure is selective for. Yet, model assumptions are always required, reducing confidence that they will generalize to empirical data, especially given the possibility that non-simulated confounding properties are present. Therefore, empirical validation is important to ensure that, at least in the cases where some ground truth can be reasonably established, the sensitivity and selectivity of a given FC measure is indeed sufficient to support mechanistic interpretation of FC observations.

Conclusions

The mechanistic FC framework developed here makes clear how many hurdles still need to be overcome to achieve full mechanistic accounts of neural network processes. Simultaneously, the framework reveals what progress we have made despite ambiguities in interpreting existing FC measures. We hope this framework catalyzes work toward improved interpretation of existing FC measures and development of FC measures (and recording techniques) that provide for more comprehensive and unambiguous inferences about brain network mechanisms of theoretical interest.

Acknowledgements

The authors would like to acknowledge the following support: National Institutes of Health R01MH107549 to LQU; National Institutes of Health R01MH109520 and R01AG055556 to MWC; National Institutes of Health P20GM103472, R01EB020407, and NSF National Science Foundation 1539067 to VC. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

Competing Financial Interests

The authors declare no competing interests.

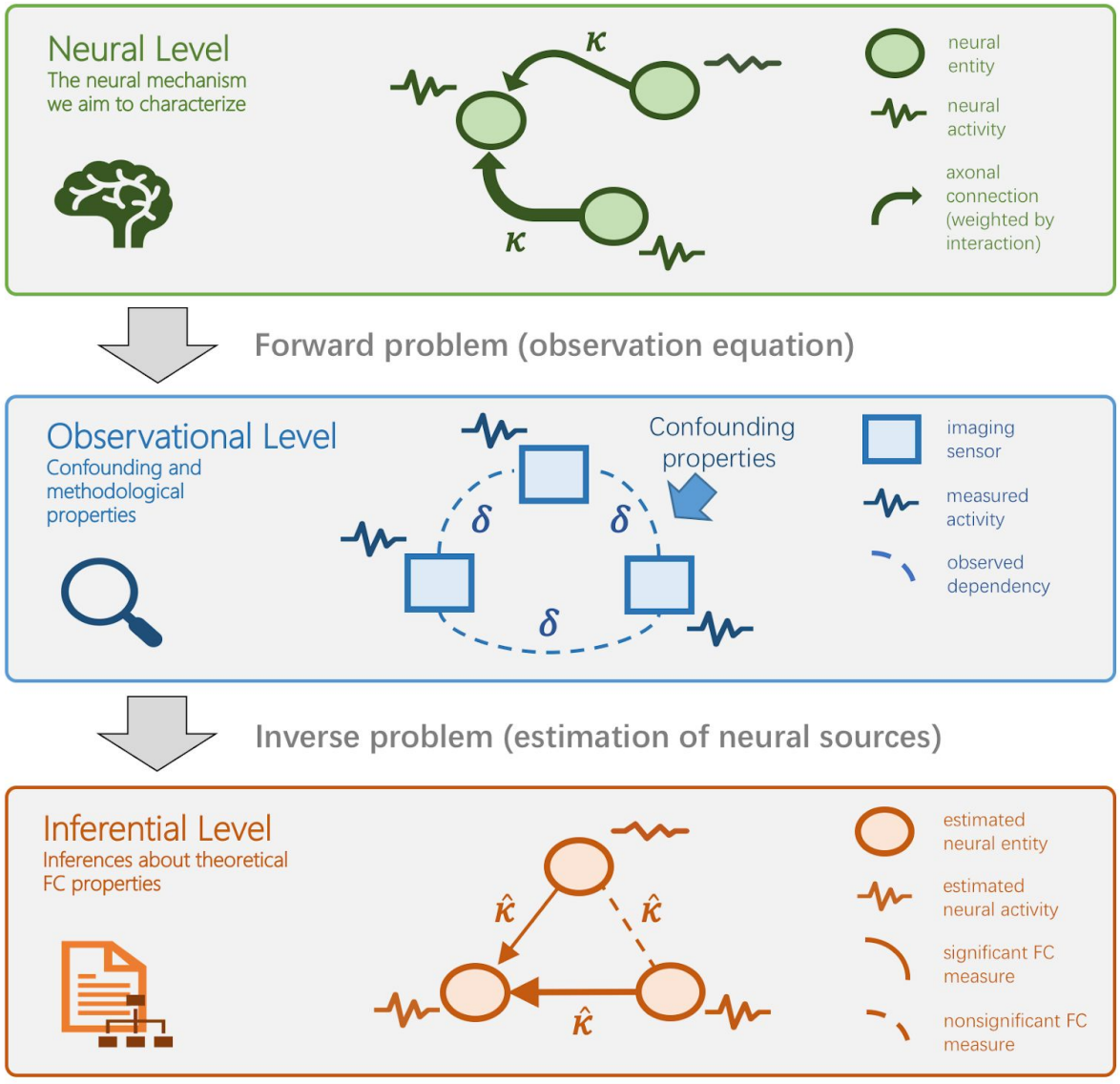

Figure 1 – Ontological levels relevant to mechanistic interpretation of FC, defining the pathway from neural mechanisms (neural level) to imaging measurements (observational level) to inferences about target theoretical properties (inferential level). At the neural level, physical connections between regions, denoted κ, depend on the signal strength (spike rate) and synaptic strength. At the observational level, time series recorded with imaging sensors (e.g., fMRI voxels, EEG electrodes, intracranial electrodes) represent neural signals that are spatiotemporally filtered through the observational pathway (forward problem). These time series also contain measurement noise and confounding variance. Observed dependencies at this level are denoted δ. At the inferential level, we attempt to infer (estimate) FC properties of interest, possibly with a degree of ambiguity, at the neural level from our observed time series. This can be done by mapping backward from sensors to neural entities (solving the inverse problem) to estimate the underlying neural activity and compute FC measures, denoted κ^hat, on this estimated activity. However, methodological and confounding properties limit the accuracy we can achieve with this backward mapping.

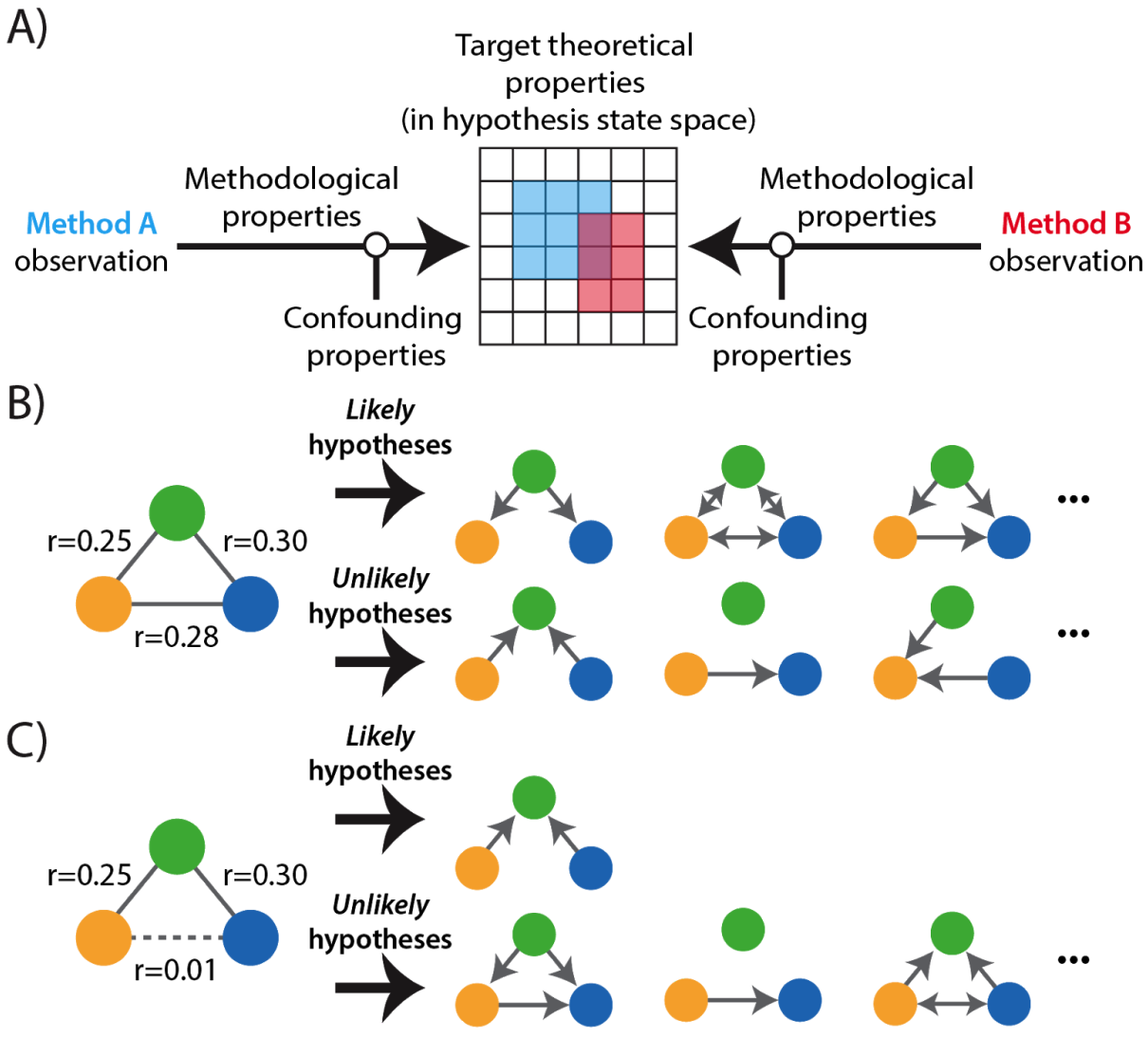

Figure 2 – The conceptual structure of the functional connectivity (FC) framework. A) Each method involves a mapping between observations and target theoretical properties, mediated by methodological properties that enable and/or restrict the possible inferences that can be made. Methodological properties merely shift what inferences are possible, but confounding properties can completely block inferences by creating ambiguities outside the space of causal brain network configurations (e.g., subject motion obscuring FC observations nullifies inferences about neural mechanisms causing FC observations). The grid illustrates the space of all hypotheses under consideration, with each grid point

being a particular causal network configuration (of which only one can be true). Each method's color indicates which hypotheses that method's results are compatible with (more coverage = more ambiguity). The overlap between methods (purple) illustrates the ability to use multiple FC methods to converge on a more narrow set of possibilities. This advances theory through logical conjunctions across FC methods. **B**) An illustration of a correlation-based FC measure in a simple 3-node network. The directionality of influences are ambiguous (based on Pearson correlation of neural time series; left side of panel) but this nonetheless constrains the hypothesis space (both likely and unlikely; right side of panel) by providing a higher probability of some causal network configurations than others. **C**) Another illustration of a simple 3-node network, this time with no correlation between the bottom two nodes. Correlation does especially well in this scenario, given that only a "collider" graph is likely with this set of correlations in a 3-node system[88].

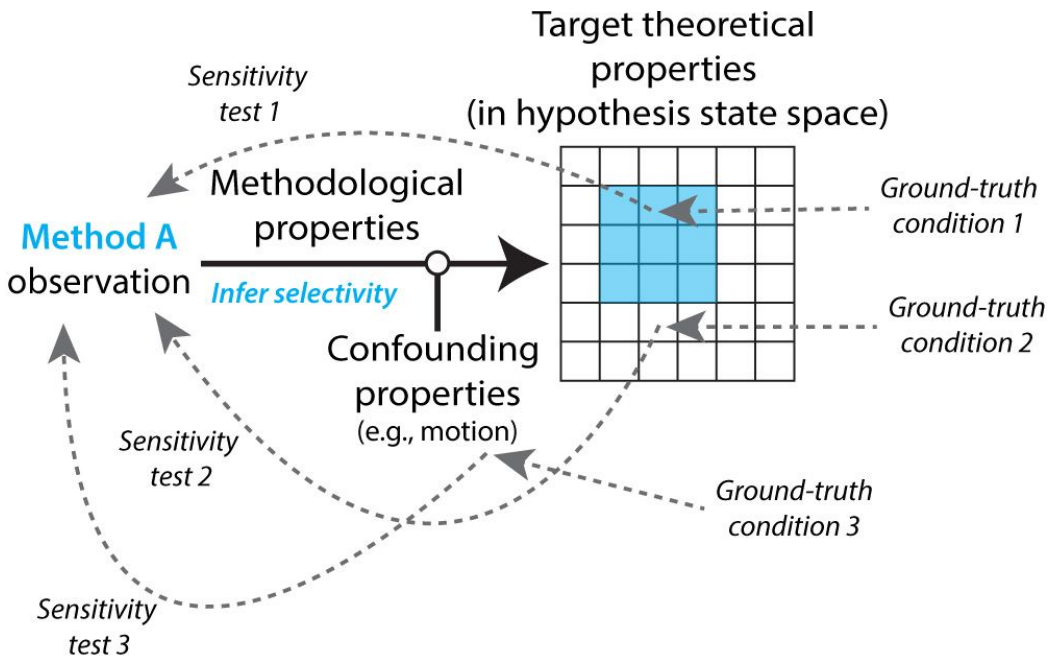

Figure 3 – A systematic approach to validate mechanistic interpretations of FC measures. The basic illustration from Figure 2 is modified with a procedure for validating FC methods. Three *ground-truth conditions* are indicated, each reflecting a scenario in which the experimenter can be highly confident about the state of the theoretical property (for the first 2 conditions) or confounding property (for the third condition) that is being manipulated. Each ground-truth condition is associated with a *sensitivity test*, wherein it is tested whether the FC method in question (Method A) is sensitive to the manipulation. Once a large set of ground-truth conditions and associated sensitivity tests has been carried out, one then infers the *selectivity* of the mapping from the Method A to theoretical and confounding properties. An FC method is valid for inferring a given set of theoretical properties of interest in so far as the method is both sensitive to and selective for those theoretical properties. The selectivity to those theoretical properties implies that the method is not sensitive to plausible confounding properties (after applying any strategies used to reduce the influence of confounds).

Table 1 - Overview of the FC framework, defining the three main types of properties relevant for drawing mechanistic inferences: theoretical, methodological, and confounding. Examples of each property are provided, along with the types of assumptions required for each. Note that these are meant to be illustrative, but not exhaustive – additional properties can be added by the researcher as appropriate.

	Theoretical properties	Methodological properties	Confounding properties
Description	Properties of the system about which the researcher would like to draw inferences. Must relate to causal interactions among neural entities.	Properties inherent to the observational or analytic methodology, which will influence the details of the inferences made regarding theoretical properties of interest	Properties of the data that may result in spurious associations that can lead to erroneous inferences regarding the theoretical properties of interest
Common Examples	• Directness: *mono- or polysynaptic* • Directionality: *A→B, A←B, A↔B* • Weight: *synaptic strength* • Linearity: *Linear or nonlinear*	• Spatial resolution/coverage • Temporal resolution • Conditions: *experimental manipulation, cohort* • Observational pathway • Neural entities: *spatially contiguous* • Interaction estimate: *e.g., correlation*	• Motion artifacts • Cardiac artifacts • Respiratory artifacts • Unmeasured neural sources • Spatial autocorrelation

Box 1. Definitions of key terms

Neural entity: *A spatially contiguous territory of neural tissue that generates a signal of interest. Examples of neural entities are: an individual neuron, a cortical column, a cortical region.*

Functional connectivity (FC): *Causal interactions between neural entities. These interactions are specified by the "theoretical properties" of causal interactions (see definition below).*

Causal interaction: *A neural event that, had it not occurred, its effect on another neural entity also would not have occurred. This is referred to as the "counterfactual" definition of causality. This definition is common in scientific reasoning: a control condition provides the alternative (counterfactual) case in which a proposed cause is altered, and observation of an altered effect constitutes evidence supporting the*

causal inference. While direct experimentally-controlled manipulation is the ideal for identifying causal interactions, decades of research suggests observational data can be used to validly constrain causal inferences in many cases.

Target theoretical property: *A property describing an aspect of causal interactions between neural entities that constitutes the inferential target of a given FC method. Examples of such properties are: directionality, directness, linearity, and weight/strength. If an FC method is not selective for a given theoretical property, it is said to be "ambiguous" with respect to that property. In Figure 1, target theoretical properties are denoted as κ, and inferences about them are denoted as κ^hat.*

Methodological property: *A property describing the observational method by which inferences about FC are supported.*

Confounding property: *A factor that induces a spurious association between the neural entities of interest, such that this association disappears if the factor is kept constant.*

Observational pathway: *An abstract description of the mapping, via representational levels, from target physical mechanisms (action potentials or postsynaptic potentials) to neural entity to sensors collecting observations. Referring to Figure 1, the observational pathway maps from real causal interactions (denoted by κ) to observed dependencies (denoted by δ). Mapping backwards from observed dependencies allows us to estimate causal interactions (denoted by κ^hat).*

FC methods: *Approaches that seek to characterize causal interactions between neural entities, with potential limitations regarding what aspects of causal inferences are valid for a given approach.*

Effective connectivity methods: *Following Friston et al.[89], approaches that use parameterized models to characterize causal interactions between neural entities. This has been operationalized as the simplest circuit diagram (parameterized model) that explains observed responses[90]. Under the proposed framework, effective connectivity methods can be considered a subset of FC methods, since both seek to characterize causal interactions between neural entities.*

Box 2. How to immediately begin improving causal inferences beyond correlation-based FC: Confound reduction using partial correlation and alternatives

Even if all confounding properties due to measurement artifacts are accounted for (e.g., motion artifacts driving spurious causal inferences) many potential confounds exist among neural entities. Such *confounders* are neural entities that directly cause activity in two or more other neural entities (Figure AA). Confounders can lead to spurious causal inferences, such as the erroneous conclusion that stimulating one neural entity (e.g., the orange node in Figure AA) will affect another neural entity (e.g., the blue node in Figure AA). This confounding problem is perhaps the biggest barrier to progress in FC research and in causal inference generally[19]. While many FC methods can make approximately correct predictions regarding effects of causal interventions despite ambiguities in other causal configurations (e.g., *chains* and *colliders*; Figure AB & AC), this is not the case for confounders[17].

The worst-case scenario for the confounding problem is when unmeasured confounders exist, given that there are limited options (e.g., directly stimulating each neural entity to observe its causal effect) for accounting for such confounders. The whole-brain coverage of modern neuroimaging methods (fMRI and EEG/MEG) provides some hope of being able to measure all neural entities at a given level of organization (e.g., brain regions). In practice, however, we are likely not observing clean signal from all neural entities of interest, given various biases in current methods (e.g., EEG/MEG signals reflecting dipoles), such that some unobserved confounders likely exist in these datasets. Yet even in the presence of unobserved confounders, taking observed confounders into account improves causal inferences. There are many FC measures that, unlike pairwise correlation and coherence, take confounding into account via fitting all measured time series simultaneously, such as partial correlation, multiple regression[91], dynamic causal modeling[40], multivariate Granger causality[92], and Bayesian search approaches[88]. We focus here on the first of these.

Partial correlation is simply the Pearson correlation between a pair of time series calculated after the portion of their variance explained by all other observed time series is removed. A partial correlation coefficient thus reflects the degree to which two time series are correlated after accounting for potential confounders represented in the other time series. This improves causal inferences in the case of confounders (Fig. AA) and chains (Fig. AB).

However, partial correlation does not improve causal inferences in the case of colliders[5,88] (Fig. AC). This is due to the regressing-out step, which ends up introducing a negative correlation (in the case of positive relationships with a collider like in Fig. AC) between independent time series. In the case illustrated in Figure AC, the orange and blue nodes' time series are mixed into the green node's time series due to causal influences (which define the green node as a collider). In this case, the regressing-out step erroneously makes two independent time series appear dependent. Note that if the orange node or blue node had been negatively related to

the green node then the regressing-out step would have introduced a positive correlation between them. These effects are related to what is sometimes referred to as “conditioning on a collider”[93].

Perhaps surprisingly, one way to correct for the confounder case is to consider the result with pairwise correlation. While pairwise correlation provides the incorrect causal adjacency graph in most cases (Fig. AA & AB), it provides the correct result in the collider case (Fig. AC). Thus, a simple approach to remove problematic partial correlation results is to remove connections that are not present with pairwise correlation but appear with partial correlation. As a further bonus, the resulting causal graph can be oriented with causal directionality (at least in a 3-node case) because only a collider graph could have produced this pattern of results[88,94,95].

Notably, this does not correct all possible problems with partial correlations in more complex graphs. For instance, consider a graph where two nodes without a direct connection between them are influenced by a confounder *and* also are themselves causes of a collider. In this graph, combining partial correlation and pairwise correlation as described above will inevitably result in a false connection between the two nodes. However, in principle a set of existing methods can account for such cases. These Bayesian search approaches combine tests of causal independence (similar to pairwise correlation) with tests for confounding (similar to partial correlation) in a causal search framework to identify the causal graph most likely to have generated the observed data[88]. Among current algorithms, we recommend the following approaches available from the Center for Causal Discovery (http://www.ccd.pitt.edu) for making causal inferences taking into account both confounders and colliders: fGES, IMaGES[2,96] (also available at https://cran.r-project.org/web/packages/IMaGES), Two-step[97], and FASK[97]. Yet even these algorithms are not suited for all conditions, such that they (like all current methods) require further refinement through theoretical and empirical validation and method development (see section “Looking forward: The central role of FC method validation”).

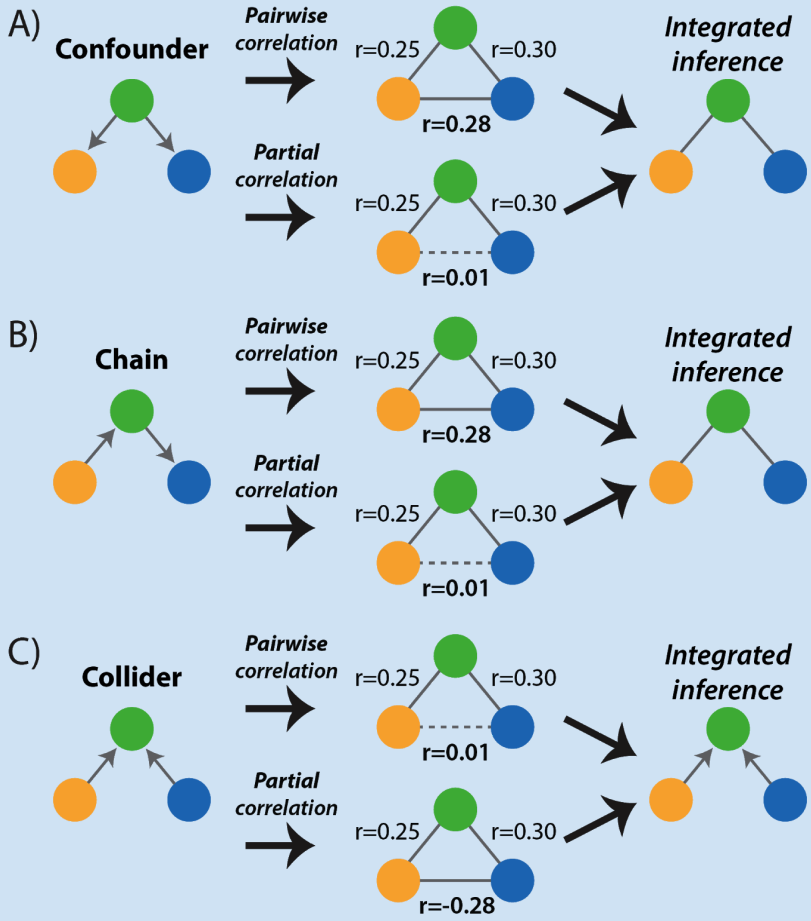

Figure A [part of Box 2] – Switching from pairwise correlation to partial correlation improves causal inference (but is not perfect). Integrating inferences from both pairwise and partial correlation improves causal inferences in most cases, though some issues remain (see text).

References

1. Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J. & Friston, K. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* **58**, 339–361 (2011).

2. Ramsey, J. D. *et al.* Six problems for causal inference from fMRI. *Neuroimage* **49**, 1545–1558 (2010).

3. Mill, R. D., Ito, T. & Cole, M. W. From connectome to cognition: The search for

mechanism in human functional brain networks. *Neuroimage* **160**, 124–139 (2017).

4. Cole, M. W., Yang, G. J., Murray, J. D., Repovš, G. & Anticevic, A. Functional connectivity change as shared signal dynamics. *J. Neurosci. Methods* **259**, 22–39 (2016).
5. Smith, S. M. The future of FMRI connectivity. *Neuroimage* **62**, 1257–1266 (2012).
6. Friston, K. J. Functional and effective connectivity: a review. *Brain Connect.* **1**, 13–36 (2011).
7. Friston, K. J. Functional and effective connectivity: a review. *Brain Connect.* **1**, 13–36 (2011).
8. Horwitz, B. The elusive concept of brain connectivity. *Neuroimage* **19**, 466–470 (2003).
9. Korzybski, A. Science and sanity: An introduction to non-aristolian systems and general semantics. (1933).
10. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
11. Power, J. D. *et al.* Functional network organization of the human brain. *Neuron* **72**, 665–678 (2011).
12. Power, J. D. & Petersen, S. E. Control-related systems in the human brain. *Curr. Opin. Neurobiol.* **23**, 223–228 (2013).
13. Smith, S. M. *et al.* Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 13040–13045 (2009).
14. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).

15. Van Dijk, K. R. A., Sabuncu, M. R. & Buckner, R. L. The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* **59**, 431–438 (2012).
16. Satterthwaite, T. D. *et al.* Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage* **60**, 623–632 (2012).
17. Mehler, D. M. A. & Kording, K. P. The lure of causal statements: Rampant mis-inference of causality in estimated connectivity. *arXiv [q-bio.NC]* (2018).
18. Pearl, J. A Probabilistic Calculus of Actions. in *Uncertainty Proceedings 1994* (eds. de Mantaras, R. L. & Poole, D.) 454–462 (Morgan Kaufmann, 1994).
19. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. (Basic Books, 2018).
20. Pearl, J., Glymour, M. & Jewell, N. P. *Causal Inference in Statistics: A Primer*. (John Wiley & Sons, 2016).
21. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512 (1986).
22. Spirtes, P., Glymour, C. & Scheines, R. Causation, prediction, and search. Adaptive computation and machine learning. (2000).
23. Marinescu, I. E., Lawlor, P. N. & Kording, K. P. Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour* **2**, 891–898 (2018).
24. Hume, D. An Enquiry concerning Human Understanding (originally published 1748). in *The Clarendon Edition of the Works of David Hume: An Enquiry concerning Human Understanding* (eds. Beauchamp, T. L., Hume, D. &

Beauchamp, T. L.) 134–198 (Oxford University Press, 2000).

25. Pearl, J., Robins, J. M. & Greenland, S. Confounding and Collapsibility in Causal Inference. *Stat. Sci.* **14**, 29–46 (1999).
26. Friston, K. J. Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78 (1994).
27. Friston, K. J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).
28. Roebroeck, A., Formisano, E. & Goebel, R. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage* **25**, 230–242 (2005).
29. Klahr, D. & Dunbar, K. Dual Space Search During Scientific Reasoning. *Cogn. Sci.* **12**, 1–48 (1988).
30. Lee, H. S., Betts, S. & Anderson, J. R. Learning Problem-Solving Rules as Search Through a Hypothesis Space. *Cogn. Sci.* **40**, 1036–1079 (2016).
31. Smith, S. M. *et al.* Network modelling methods for FMRI. *Neuroimage* **54**, 875–891 (2011).
32. Mill, R. D., Bagic, A., Bostan, A., Schneider, W. & Cole, M. W. Empirical validation of directed functional connectivity. *Neuroimage* **146**, 275–287 (2017).
33. Wang, H. E. *et al.* A systematic framework for functional connectivity measures. *Front. Neurosci.* **8**, 405 (2014).
34. Illari, P. M. & Williamson, J. What is a mechanism? Thinking about mechanisms across the sciences. *Eur. J. Philos. Sci.* **2**, 119–135 (2012).
35. Hutchison, R. M. *et al.* Dynamic functional connectivity: Promises, issues, and interpretations. *Neuroimage* **80**, 360–378 (2013).
36. Lurie, D. *et al.* On the nature of resting fMRI and time-varying functional

connectivity. *PsyArXiv Preprints* (2018).

37. Smith, S. M. *et al.* A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18**, 1565–1567 (2015).
38. Schultz, D. H. & Cole, M. W. Higher Intelligence Is Associated with Less Task-Related Brain Network Reconfiguration. *J. Neurosci.* **36**, 8551–8561 (2016).
39. Cole, M. W., Yarkoni, T., Repovs, G., Anticevic, A. & Braver, T. S. Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *J. Neurosci.* **32**, 8988–8999 (2012).
40. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
41. Frässle, S. *et al.* A generative model of whole-brain effective connectivity. *Neuroimage* **179**, 505–529 (2018).
42. Honey, C. J., Kötter, R., Breakspear, M. & Sporns, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10240–10245 (2007).
43. Lohmann, G., Erfurth, K., Müller, K. & Turner, R. Critical comments on dynamic causal modelling. *Neuroimage* **59**, 2322–2329 (2012).
44. Lewontin, R. C. & Others. *The genetic basis of evolutionary change*. **560**, (Columbia University Press New York, 1974).
45. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).
46. Hines, M. L. & Carnevale, N. T. The NEURON simulation environment. *Neuron* **9**, (2006).

47. Goodman, D. Brian: a simulator for spiking neural networks in Python. *Front. Neuroinform.* **2**, (2008).
48. Ramsey, J. D., Hanson, S. J. & Glymour, C. Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simulation study. *Neuroimage* **58**, 838–848 (2011).
49. Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F. & Danks, D. Causal discovery from subsampled time series data by constraint optimization. (2016).
50. Schubert, N. *et al.* 3D Reconstructed Cyto-, Muscarinic M2 Receptor, and Fiber Architecture of the Rat Brain Registered to the Waxholm Space Atlas. *Front. Neuroanat.* **10**, 51 (2016).
51. Eickhoff, S. B. *et al.* A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* **25**, 1325–1335 (2005).
52. Craddock, R. C., James, G. A., Holtzheimer, P. E., 3rd, Hu, X. P. & Mayberg, H. S. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* **33**, 1914–1928 (2012).
53. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).
54. Potjans, T. C. & Diesmann, M. The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cereb. Cortex* **24**, 785–806 (2014).
55. Logothetis, N. K. The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357**, 1003–1037 (2002).

56. Buxton, R. B., Wong, E. C. & Frank, L. R. Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magn. Reson. Med.* **39**, 855–864 (1998).

57. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).

58. Birn, R. M., Saad, Z. S. & Bandettini, P. A. Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. *Neuroimage* **14**, 817–826 (2001).

59. Jellema, W. T. *et al.* Heterogeneity and prediction of hemodynamic responses to dobutamine in patients with septic shock. *Crit. Care Med.* **34**, 2392–2398 (2006).

60. Tarantini, S., Tran, C. H. T., Gordon, G. R., Ungvari, Z. & Csiszar, A. Impaired neurovascular coupling in aging and Alzheimer's disease: Contribution of astrocyte dysfunction and endothelial impairment to cognitive decline. *Exp. Gerontol.* **94**, 52–58 (2017).

61. Handwerker, D., Ollinger, J. & D'Esposito, M. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* **21**, 1639–1651 (2004).

62. Di, X., Kannurpatti, S. S., Rypma, B. & Biswal, B. B. Calibrating BOLD fMRI activations with neurovascular and anatomical constraints. *Cereb. Cortex* **23**, 255–263 (2013).

63. Rangaprakash, D., Wu, G.-R., Marinazzo, D., Hu, X. & Deshpande, G. Hemodynamic response function (HRF) variability confounds resting-state fMRI functional connectivity. *Magn. Reson. Med.* (2018). doi:10.1002/mrm.27146

64. Calhoun, V. D., Stevens, M. C., Pearlson, G. D. & Kiehl, K. A. fMRI analysis with the general linear model: removal of latency-induced amplitude bias by

incorporation of hemodynamic derivative terms. *Neuroimage* **22**, 252–257 (2004).

65. Pearl, J., Robins, J. M. & Greenland, S. Confounding and Collapsibility in Causal Inference. *Stat. Sci.* **14**, 29–46 (1999).
66. Schoffelen, J.-M. & Gross, J. Source connectivity analysis with MEG and EEG. *Hum. Brain Mapp.* **30**, 1857–1865 (2009).
67. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449–468 (2014).
68. Sochat, V. *et al.* A robust classifier to distinguish noise from fMRI independent components. *PLoS One* **9**, e95493 (2014).
69. Acharjee, P. P., Phlypo, R., Wu, L., Calhoun, V. D. & Adali, T. Independent Vector Analysis for Gradient Artifact Removal in Concurrent EEG-fMRI Data. *IEEE Trans. Biomed. Eng.* **62**, 1750–1758 (2015).
70. Du, Y. *et al.* Artifact removal in the context of group ICA: A comparison of single-subject and group approaches. *Hum. Brain Mapp.* **37**, 1005–1025 (2016).
71. Glasser, M. F. *et al.* Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. *Neuroimage* **181**, 692–717 (2018).
72. Buibas, M. & Silva, G. A. A framework for simulating and estimating the state and functional topology of complex dynamic geometric networks. *Neural Comput.* **23**, 183–214 (2011).
73. Fuentes, L., Aldana, J. F. & Troya, J. M. GENESIS: An Object-Oriented Framework for Simulation of Neural Network Models. in *Artificial Neural Nets and Genetic Algorithms* (eds. Pearson, D. W., Steele, N. C. & Albrecht, R. F.) 321–324 (Springer Vienna, 1995).

74. Ritter, P., Schirner, M., McIntosh, A. R. & Jirsa, V. K. The Virtual Brain Integrates Computational Modeling and Multimodal Neuroimaging. *Brain Connect.* **3**, 121–145 (2013).

75. Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M. & Friston, K. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* **4**, e1000092 (2008).

76. David, O., Cosmelli, D. & Friston, K. J. Evaluation of different measures of functional connectivity using a neural mass model. *Neuroimage* **21**, 659–673 (2004).

77. Gourévitch, B., Bouquin-Jeannès, R. L. & Faucon, G. Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biol. Cybern.* **95**, 349–369 (2006).

78. Wang, Y., Katwal, S., Rogers, B., Gore, J. & Deshpande, G. Experimental Validation of Dynamic Granger Causality for Inferring Stimulus-Evoked Sub-100 ms Timing Differences from fMRI. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 539–546 (2017).

79. Nee, D. E. & D'Esposito, M. Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *Elife* **6**, (2017).

80. Wheeler, M., Petersen, S. & Buckner, R. Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 11125–11129 (2000).

81. David, O. *et al.* Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* **6**, 2683–2697 (2008).

82. Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J. & Jarvis, E. D. Computational

inference of neural information flow networks. *PLoS Comput. Biol.* **2**, e161 (2006).

83. Ryali, S. *et al.* Combining optogenetic stimulation and fMRI to validate a multivariate dynamical systems model for estimating causal brain interactions. *Neuroimage* **132**, 398–405 (2016).
84. Lee, J. H. Informing brain connectivity with optogenetic functional magnetic resonance imaging. *Neuroimage* **62**, 2244–2249 (2012).
85. Power, J. D., Schlaggar, B. L. & Petersen, S. E. Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* **105**, 536–551 (2015).
86. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320–341 (2014).
87. Cole, M. W. *et al.* Task activations produce spurious but systematic inflation of task functional connectivity estimates. *Neuroimage* **189**, 1–18 (2019).
88. Mumford, J. A. & Ramsey, J. D. Bayesian networks for fMRI: A primer. *Neuroimage* **86**, 573–582 (2014).
89. Friston, K., Moran, R. & Seth, A. K. Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* **23**, 172–178 (2013).
90. Aertsen, A. M., Gerstein, G. L., Habib, M. K. & Palm, G. Dynamics of neuronal firing correlation: modulation of‘ effective connectivity’. *J. Neurophysiol.* **61**, 900–917 (1989).
91. Cole, M. W., Ito, T., Bassett, D. S. & Schultz, D. H. Activity flow over resting-state networks shapes cognitive task activations. *Nat. Neurosci.* **19**, 1718–1726 (2016).
92. Barnett, L. & Seth, A. K. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *J. Neurosci. Methods* **223**, 50–68 (2014).
93. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer New York,

2016).

94. Rebane, G. & Pearl, J. The Recovery of Causal Poly-Trees from Statistical Data. *Proceedings of the Third Workshop on Uncertainty in AI* 222–228 (1987).
95. Schiefer, J. *et al.* From correlation to causation: Estimating effective connectivity from zero-lag covariances of brain signals. *PLoS Comput. Biol.* **14**, e1006056 (2018).
96. Ramsey, J., Glymour, M., Sanchez-Romero, R. & Glymour, C. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* **3**, 121–129 (2017).
97. Sanchez-Romero, R. *et al.* Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Netw Neurosci* **3**, 274–306 (2019).