Title

Causally informed activity flow models provide mechanistic insight into the emergence of cognitive processes from brain network interactions

Authors

Ruben Sanchez-Romero[1, *], Takuya Ito[1, 3], Ravi D. Mill[1], Stephen José Hanson[2], Michael W. Cole[1]

[1] Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ, 07102, USA
[2] Rutgers University Brain Imaging Center (RUBIC), Rutgers University, Newark, NJ, 07102, USA
[3] Current affiliation: Department of Psychiatry, Yale University School of Medicine, New Haven, CT, 06510, USA
*Corresponding author: ruben.saro@rutgers.edu (R. Sanchez-Romero)

Format: Original research paper

Highlights

- Activity flow models provide insight into how cognitive neural effects emerge from brain network interactions.
- Functional connectivity methods grounded in causal principles facilitate mechanistic interpretations of task activity flow models.
- Mechanistic activity flow models accurately predict task-evoked neural effects across a wide variety of brain regions and cognitive tasks.

Abstract

Brain activity flow models estimate the movement of task-evoked activity over brain connections to help explain the emergence of task-related functionality. Activity flow estimates have been shown to accurately predict task-evoked brain activations across a wide variety of brain regions and task conditions. However, these predictions have had limited explanatory power, given known issues with causal interpretations of the standard functional connectivity measures used to parameterize activity flow models. We show here that functional/effective connectivity (FC) measures grounded in causal principles facilitate mechanistic interpretation of activity flow models. Starting from Pearson correlation (the current field standard), we progress from FC measures with poor to excellent causal grounding, demonstrating a continuum of causal validity using simulations and empirical fMRI data. Finally, we apply a causal FC method to a dorsolateral prefrontal cortex region, demonstrating causal network mechanisms contributing to its strong activation during a 2-back (relative to a 0-back) working memory task. Together, these results reveal the promise of parameterizing activity flow models using causal FC methods to identify network mechanisms underlying cognitive computations in the human brain.

Keywords: predictive models; causal inference; brain networks; functional connectivity; network neuroscience; activity flow

Data/code availability statement: Data and code to reproduce our analyses will be publicly available at the project repository upon manuscript acceptance. (This information is stated in Materials and Methods.)

Ethics statement: We used open access data from the Human Connectome Project (HCP), for which all subjects gave signed informed consent in accordance with the protocol approved by the Washington University institutional review board. We abide by the HCP open access use terms and the Rutgers University institutional review board approved use of these data (This information is included in Materials and Methods.)

Declaration of interests: The authors have no conflict of interest to declare.

Author statement:
Ruben Sanchez-Romero: Conceptualization, Methodology, Software, Formal Analysis, Writing-original draft, Writing-review & editing, Visualization. **Takuya Ito**: Conceptualization, Methodology, Investigation, Data Curation, Writing-review & editing. **Ravi D. Mill**: Conceptualization, Methodology, Investigation, Data Curation, Writing-review & editing. **Stephen José Hanson**: Conceptualization, Methodology, Resources, Writing-review & editing, Funding acquisition. **Michael W. Cole**: Conceptualization, Methodology, Software, Investigation, Resources, Writing-original draft, Writing-review & editing, Supervision, Funding acquisition.

Introduction

The activity flow mapping (actflow) framework (Cole et al., 2016) aims to model the emergence of observed neural responses from functional interactions between neural populations. Actflow is built upon a long tradition of artificial neural networks, spanning highly abstract to strongly biologically-constrained models (Fukushima, 1980; Stephen José Hanson & Burr, 1990; Hopfield, 1982; Kohonen, 1984; McClelland & Rogers, 2003; McCulloch & Pitts, 1943; Olson & Hanson, 1990; Rashevsky, 1933; Rosenblatt, 1958; Rumelhart, Hinton, & Williams, 1986; Von der Malsburg, 1973; Widrow, 1962; Yamins et al., 2014). The actflow framework has successfully imported insights from neural networks (in particular, the propagation and activation rules (Rumelhart, Hinton, & McClelland, 1986)) and applied them to research the complex interplay between functional connectivity and task-evoked neural responses. Specifically, actflow has been used to study: the flow of task-related activity via whole-brain resting-state networks (Cole et al., 2016); the fine-scale transfer of information-representing task activity between specific pairs of functionally connected regions (Ito et al., 2017); the relevance of task-state functional networks in communicating task-related neural responses (Cole et al., 2021); the disruption of task activations from altered functional networks in pre-clinical Alzheimer's disease (Mill et al., 2020); the disruption of task activations from altered activity flows in schizophrenia (Hearne et al., 2020); the role of specific brain networks in a visual shape completion task (Keane et al., 2020); and the cortical heterogeneity of localized and distributed cognitive processes (Ito, Hearne, & Cole, 2020).

In general, cognitive tasks are experimental manipulations designed to cause a series of neural responses (activations), which can then be focally estimated (e.g., using regression). These estimated neural activations can be considered causal effects of an exogenous experimental intervention (Maathuis et al., 2009), but by themselves do not explain the underlying causal mechanisms from which they emerge. To bridge this gap the actflow framework provides a connectionist model of the generation of task-related activation that combines functional connectivity (FC) patterns and neural activations propagating through those connections (**Figure 1G**). The studies mentioned above confirm that actflow modeling can indeed successfully predict effects of task experimental interventions, while providing hypotheses about the causal network interactions that give rise to such effects.

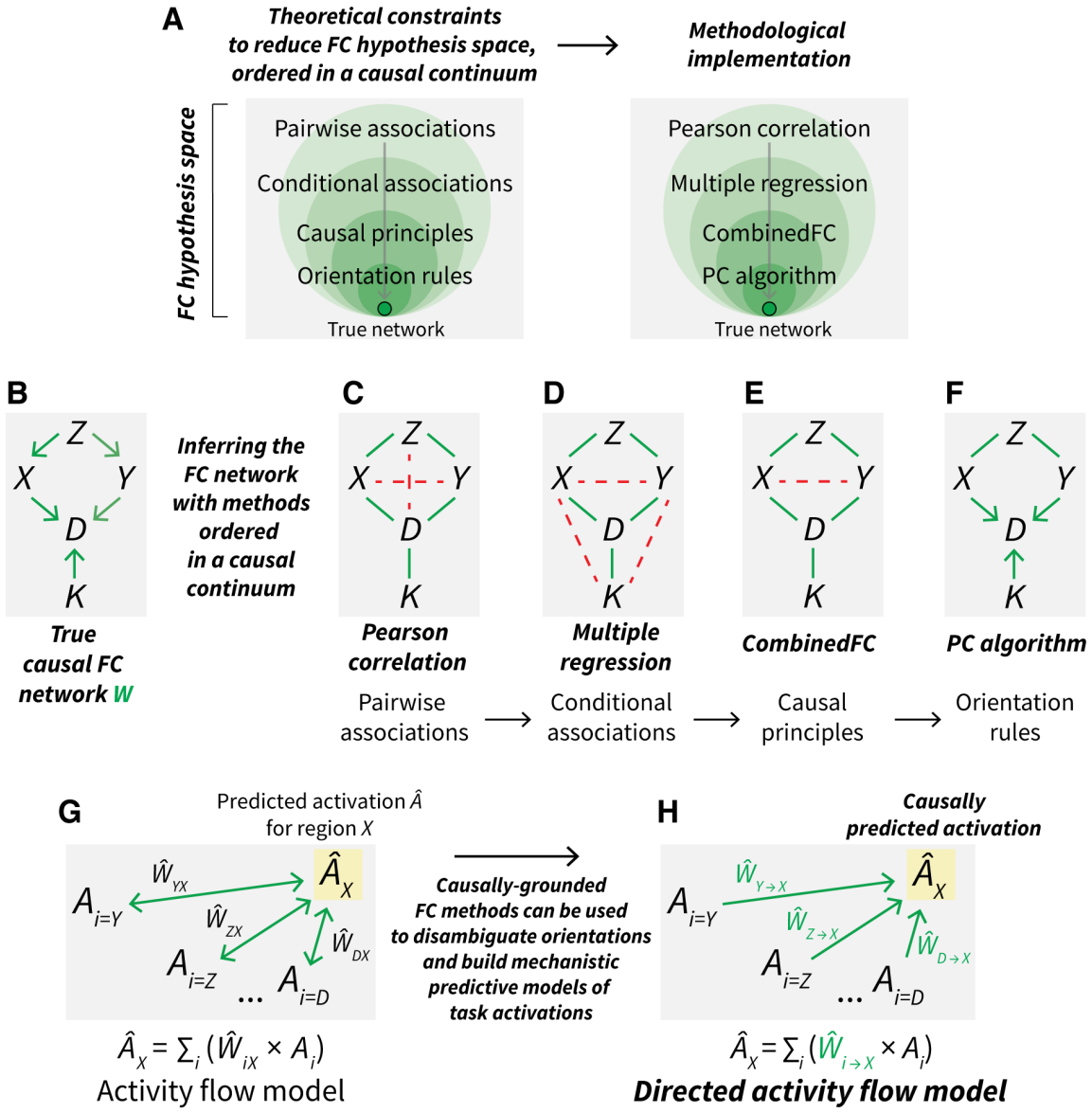

Figure 1 — Causally-grounded functional connectivity methods can be used to build mechanistic activity flow predictive models. (**A**) When inferring an FC network from neural data, theoretical constraints ordered in a causal continuum can be used to reduce the possibly vast hypothesis space. These constraints can build upon each other in order to get closer to the true causal network. Here we implemented them with four different FC methods: correlation, multiple regression, combinedFC and PC algorithm. (**B**) An example of a true causal FC network, denoted as *W*, for five neural regions time series *Z*, *X*, *Y*, *D* and *K*. The green arrows represent direct causal functional associations between the time series. (**C**) The expected network when using Pearson correlation FC to recover the true mechanism from panel B. Correlation only evaluates pairwise associations between time series. Green lines indicate correctly inferred undirected connections, while red dashed lines indicate incorrectly inferred connections. Incorrectly inferred connections resulted from not controlling for one causal confounder ($X \leftarrow Z \rightarrow Y$) and two causal chains ($Z \rightarrow X \rightarrow D$ and $Z \rightarrow Y \rightarrow D$). (**D**) The expected network when using multiple regression FC. The association between each pair of time series is conditioned on the rest of the regions to control for confounders and chains. In this case, the incorrectly inferred connections resulted from three conditioned-on causal colliders ($X \rightarrow D \leftarrow Y$, $X \rightarrow D \leftarrow K$ and $Y \rightarrow D \leftarrow K$). (**E**) The network

recovered by combinedFC. Thanks to its zero-correlation check—based on the statistical behavior of colliders (see CombinedFC section)—combinedFC removed two of the spurious connections from conditioned-on colliders. Nevertheless, for this mechanism the zero-correlation check cannot remove the remaining spurious connection because the confounder ($X \leftarrow Z \rightarrow Y$) will always force a non-zero correlation between X and Y. (**F**) The directed network recovered by the PC algorithm. By iteratively testing associations with conditioning sets of increasing size and then applying a series of orientation rules, PC inferred the true FC mechanism from panel C, with the exception of the direction of two connections (see the PC algorithm section for details on why some connections cannot be oriented by this method). (**G**) Activity flow predictive model. The inferred task-evoked activation $\hat{A}$ of a held-out neural region X (yellow box) is predicted as a linear function of the inferred FC weights ($\hat{W}$, green bidirectional arrows) and the actual task-evoked activations A from the rest of the i connected regions. The bidirectional arrows reflect the ambiguity of the inferred FC with respect to the true causal orientation. (**H**) Directed activity flow model using data-driven information about the causal direction of the functional connections. The task-evoked activation $\hat{A}$ of a neural region X (yellow box) is predicted as a linear function of the causal FC weights ($\hat{W}$, green unidirectional arrows) and the actual task-evoked activations A from the i causal source regions. The unidirectional arrows show that causally-valid FC methods can be helpful to infer the causal direction of connections.

Actflow models parameterized with functional networks estimated with Pearson correlation (pairwise association; the field standard) or with multiple regression (fully conditional associations) have provided accurate predictions of task-evoked activations. This suggests that these methods can capture, to some degree, relevant properties of the mechanisms supporting task-related functionality. Nevertheless, these FC methods pose theoretical limitations for causal interpretation of actflow models. For example, they are inherently undirected, and from their results we cannot make inferences about the causal direction of the activity flow evoked by a task manipulation. In addition, these FC methods are known to have issues with spurious connections arising from uncontrolled common causes (causal confounders, $X \leftarrow Z \rightarrow Y$), uncontrolled causal chains ($X \rightarrow Z \rightarrow Y$), and incorrectly controlled common effects (causal colliders, $X \rightarrow Z \leftarrow Y$) (Reid et al., 2019). These spurious inferences can bias actflow predictions by producing incorrect hypotheses of the true underlying functional networks supporting task computations.

To overcome these limitations, we propose building more causally valid actflow models parameterized with directed FC networks, which estimate the task-evoked activation of a particular neural region (causal effect of a task manipulation) as a function of the activation values of its direct causal source regions (**Figure 1H**). With these models, we are not only interested in accurate predictions of task activity but also in identifying causal networks contributing to the emergence of task functionality. FC methods grounded in stronger causal principles about the data generating mechanism, such as combinedFC (Sanchez-Romero & Cole, 2021) and methods based on Bayes networks (Mumford & Ramsey, 2014), allow for more effective control over confounders, chains and colliders than standard correlation and regression approaches. We hypothesize that these methods will produce FC networks whose connections more accurately (relative to standard FC estimates) represent true direct interactions between regions. This will improve the theoretical insight about neurocognitive function generated by actflow predictions by providing more valid mechanistic accounts of the functional interactions between neural populations that underlie task-related computations.

The ability to infer causal directions and avoid spurious connections using causal principles suggests a continuum into which we can order FC methods (**Figure 1A**). Here, we test the field-standard Pearson correlation (located at the least causal extreme of the continuum), followed by multiple regression (progressing one degree in the continuum, thanks to the use of conditional associations), then combinedFC (progressing another degree due to the use of causal principles to detect the presence of causal colliders), and finishing with the Bayes networks-based PC (Peter-Clark) algorithm (Spirtes et al., 2000) (progressing one more degree by integrating selective conditional associations and rules to orient connections) (**Figures 1B-F**). We use the PC algorithm—which, like all current methods, is imperfect—because we hypothesized that it provides much more accurate causal inferences than the current field standard in FC research (Pearson correlation). Further, we consider the PC algorithm to be an important and tractable example of how to use causal principles to effectively integrate pairwise and conditional association methods, as well as how causal directionality can be derived from these principles (see PC algorithm section and **Figure 1F**).

To test our hypothesis about the varying causal validity of these FC methods (**Figure 1A**), we first use simulated resting-state functional MRI (fMRI) data to determine the methods' accuracy in recovering connectivity patterns of ground-truth (since we specified the simulation parameters ourselves) FC networks. Our subsequent hypothesis is that more causally-valid FC methods, by better controlling the effects of spurious connections, will lead to actflow models with better predictions of task-evoked activations. To test this second hypothesis, we simulate task-state fMRI data and measure the prediction efficacy of actflow models based on these different FC methods.

Theoretical considerations about these FC methods' causal validity, and their performance on simulated fMRI data, prompt us to hypothesize that the observed comparative performance on simulations will translate, to a degree, to empirical data. To test this hypothesis, we use empirical fMRI data to measure the accuracy of actflow models in predicting empirical task-evoked activations, across a large battery of cortical regions, task conditions and participants.

The above general analysis motivates us to show at more detail the differences in predictive efficacy between activity flow models based on the field-standard Pearson correlation FC, and those based on the PC algorithm FC (the method with the strongest causal principles in our proposed causal validity continuum). To do this, we compare the methods in how well they can predict whole-brain task activation patterns for each of the task conditions analyzed, and how well they can predict for each of the individual brain regions, the activations evoked by the whole set of task conditions.

Finally, we illustrate how actflow models parameterized with directed FC methods can provide mechanistic insight into the emergence of causal effects from task experimental interventions in specific brain regions. In particular, we applied the PC algorithm to a right dorsolateral prefrontal cortex region to identify likely causal networks contributing to an activation contrast between a 2-back and a 0-back condition of the n-back working memory task. This local actflow

model characterizes the flow of task-related activity from *direct* causal source regions to a dorsolateral prefrontal target region. As such, its mechanistic insight represents a starting point for what could be considered a full neurocognitive mechanistic explanation, which would identify the full causal chain from stimulus to network-supported cognitive neural effects to behavioral response (Ito, Hearne, Mill, et al., 2020; Weichwald & Peters, 2021).

Confirmation that some FC methods are more causally valid and lead to better actflow predictions of empirical data would demonstrate the validity of using actflow models to develop plausible (i.e., above-chance/non-random) causal explanations for the emergence of cognitive effects (task-evoked activations) from brain data.

Materials and Methods

Activity flow mapping

Activity flow mapping (actflow) is a general predictive model to explain local task-related neural activations as the product of task-evoked activity flowing through pathways of functional brain connections (Cole et al., 2016). Formally, for a set of brain regions **V**, the task-related activation A_X for brain region X, can be expressed as $A_X = f(\mathrm{W}_X, \mathrm{A}_{\mathbf{V}\setminus\{X\}})$, where W_X are the connections of X with the rest of the regions, $\mathrm{A}_{\mathbf{V}\setminus\{X\}}$ are the activations of all regions in **V** except X, and $f()$ is a function relating connections and activations. Following Cole et al. (2016), we assume $f()$ is a linear function and implement the actflow prediction model for a particular held-out region as $\hat{A}_X = \Sigma_{i\in\mathbf{V}\setminus\{X\}}\hat{W}_{iX}A_i$, where the predicted activation ($\hat{A}_X$) is the sum of the actual activations of all other regions (A_i), weighted by their estimated connectivity values with X ($\hat{W}_{iX}$) (**Figure 1G**). (This function corresponds to a neural network linear propagation and activation rule (Stephen José Hanson & Burr, 1990; Rumelhart, Hinton, & McClelland, 1986).)

The above definition of actflow does not differentiate between causal sources and causal targets in the connectivity pattern, and predicts using all inferred connected regions. By using all truly connected regions to predict the activation of a particular held-out region we are leveraging information from the direct causes (as the linear combination of the afferent (incoming) connection weights and the sources activations), and from the direct effects (as the linear combination of the efferent (outgoing) connection weights and the targets activations). Once we consider the activation of the direct causes and of the direct effects of a held-out region, information from other regions loses its predictive power, since all relevant information to predict the held-out region is already accounted for (Guyon et al., 2008). Thus, we can predict with high accuracy the activation of a held-out region only using its truly connected regions, disregarding if these are direct sources or direct targets. This implies that an actflow model can be parameterized with an undirected functional connectivity method and achieve a high prediction accuracy, as long as the method effectively avoids spurious connections (Cole et al., 2016).

To test the improvements on actflow predictions derived from stronger causal principles in the FC methods, we used three undirected association methods on fMRI time series: correlation, multiple regression and combinedFC.

Despite the likelihood that using all connected regions is optimal for predicting task-evoked activations, we want to impose a causal biological constraint on the actflow models, such that the activation of a region is modelled as a function of the task activity flowing into it from its direct causal sources (**Figure 1H**). Formally, the prediction of the activity of a held-out region must be derived from the actual task-evoked activation of its direct sources and the corresponding afferent connection weights. From a mechanistic perspective, our goal now is not just to maximize the accuracy of our prediction, but to correctly predict how the activation of a held-out region will react to exogenous or endogenous changes in its direct causes, its afferent connection weights, or both. Nevertheless, moving towards this kind of mechanistic predictive model comes with a potential reduction of predictive power, since we will predict a held-out region only using its direct causes, and not leveraging useful information from its direct effects (if any).

With this idea in mind, we define a mechanistic linear actflow model as $\hat{A}_X = \Sigma_{i \in \mathbf{v} \setminus \{X\}} \hat{W}_{i \to X} A_i$, where $\hat{W}_{i \to X}$ are the estimated causal connections from direct sources i to held-out region X. In contrast to the first actflow definition (**Figure 1G**), this causal model predicts the task-evoked activation for a target region using *only* its estimated causal sources (**Figure 1H**). The challenge with this mechanistic model is that to obtain connectivity estimates $\hat{W}_{i \to X}$, we necessarily need a directed FC method. Here, we use the PC algorithm to estimate the required causal directed networks.

Finally, as in Cole et al. (2016), we measured the prediction accuracy of actflow models using the Pearson correlation r between predicted and actual activations, and compared it across the different FC methods used to parameterize the models. Activity flow mapping prediction and evaluation analyses were performed with the Python open-source Actflow Toolbox (available at colelab.github.io/ActflowToolbox).

Functional connectivity methods

Correlation

Functional connectivity methods can be organized in a continuum depending on their causal principles. On the less-causal extreme we place pairwise associative methods, such as Pearson correlation or mutual information (a way to measure non-linear statistical associations). These methods do not hold causal assumptions about the generating mechanism giving rise to the observed association. For example, a non-zero Pearson correlation between the time series of two brain regions X and Y indicates a functional association between these regions, but no further knowledge about the nature of this association can be derived from it. We cannot conclude if the observed non-zero pairwise correlation resulted from a causal mechanism where one brain region is the direct cause of the other ($X \to Y$ or $X \leftarrow Y$), or one region is the indirect cause of the other (causal chain, $X \to Z \to Y$), or a third region is a common cause of the two

regions (causal confounder, $X \leftarrow Z \rightarrow Y$) (Reichenbach, 1956), or a combination of these cases. This ambiguity impedes causal interpretations of correlation-based network connections. In practice, pairwise associative methods do not control for the effects of causal confounders and chains, thus inferring spurious connections in the estimated FC network (**Figure 1C**). In activity flow models, these spurious connections will create false pathways through which task activity will be incorrectly added, biasing the linear prediction of the task-evoked activations.

We applied Pearson correlation $r_{XY} = cov(X,Y)/(std(X)std(Y))$, where X and Y are time series for two brain regions, $cov()$ is the sample covariance and $std()$ is the sample standard deviation; and used a two-sided z-test with significance threshold of p-value $< \alpha = 0.01$, for every individual simulation and empirical dataset.

Multiple regression

Advancing on the causal continuum, we can use multiple regression to compute the statistical association between one region time series and every member of a set of regressor time series—in FC analysis this set is usually the rest of the brain regions in the dataset—where each association is conditioned on the rest of the regressors. Conditioning on the rest of the regions in the dataset controls for spurious connections arising from the effect of causal chains and causal confounders. Thus, in contrast to pairwise methods, a non-zero multiple regression coefficient can be interpreted as a direct functional connection between two brain regions. Despite this improvement in causal interpretation, multiple regression still cannot determine the actual causal orientation of direct network connections. In addition, a fundamental limitation of multiple regression as an FC method is that by conditioning on the rest of the brain regions it will infer a spurious association between two unconnected regions if these two regions are causes of a third one (collider, $X \rightarrow Z \leftarrow Y$) (**Figure 1D**) (Berkson, 1946; Bishop, 2006; Kiiveri et al., 1984; Reid et al., 2019). The presence of colliders in a causal structure—which we cannot tell in advance—implies that any connection inferred by multiple regression could in principle be a spurious connection. For example, Sanchez-Romero & Cole (2021) showed that in simulated networks with a larger proportion of colliders relative to confounders, multiple regression returns a higher number of spurious connections than correlation.

We applied ordinary least squares linear multiple regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e_Y$, where Y is the time series for a brain region, X_1 to X_k are the time series for the rest of the regions in the dataset, β_1 to β_k are the corresponding regression coefficients, β_0 is the intercept and e_Y is the regression error of Y. We used a two-sided t-test for the regression coefficients with significance threshold of p-value $< \alpha = 0.01$, for every individual simulation and empirical dataset.

CombinedFC

The combined functional connectivity (combinedFC) method (Sanchez-Romero & Cole, 2021), proposes a causally-principled solution to avoid spurious connections from conditioning on colliders. As such, combinedFC allows us to advance in the causal continuum from correlation and multiple regression to more causally-valid methods of functional connectivity. The strategy of combinedFC is based on the observation that for a collider $X \rightarrow Z \leftarrow Y$, the pairwise

correlation of the two causes *X* and *Y* will be *zero*; while the multiple regression $X = aY + bZ$ (or $Y = aX + bZ$), where the common effect *Z* is being conditioned on, will infer a *non-zero* regression coefficient *a* between *X* and *Y*.

CombinedFC leverages this observation to detect and remove spurious connections from conditioning on colliders. In a first step, the method computes the multiple regression for each brain region on the rest of the regions in the dataset. In a second step, it checks for each *non-zero* multiple regression coefficient, if its corresponding pairwise correlation is *zero*. If this is the case, there is evidence of a spurious connection from conditioning on a collider and combinedFC removes the false connection from the network (**Figure 1E**).

Using combinedFC we have more evidence to conclude that the inferred network *does not* include indirect causal chains or spurious connections from causal confounders—thanks to the initial multiple regression conditioning—or spurious connections from conditioning on colliders—thanks to the zero-correlation check. Nonetheless, Sanchez-Romero & Cole (2021) have shown that in the presence of certain challenging causal patterns, for example a mix of a causal confounder and a collider (e.g., **Figure 1B**), combinedFC will inevitably produce spurious connections (**Figure 1E**). It is the risk of these spurious connections that limits combinedFC to unequivocally orient detected colliders. Thus, as correlation and multiple regression, combinedFC returns an undirected network, albeit with more causally-interpretable connections.

Here, we implement combinedFC with two modifications. In the first step of combinedFC, instead of using linear multiple regression to evaluate conditional associations, we used partial correlation using the inverse of the covariance matrix, which is a faster and equivalent way to determine significant connections (Sanchez-Romero & Cole, 2021). The second modification is considering the output of combinedFC as an initial feature selection step (Guyon et al., 2008; Guyon & Elisseeff, 2003), and compute the final FC weights by regressing each region only on its connected regions (selected features) in the combinedFC network. We have seen that this second modification produces FC weights that result in higher activity flow prediction accuracy than the original weights of combinedFC.

For every individual simulation and empirical dataset, we applied combinedFC with a two-sided *z*-test and significance threshold of $p\text{-value} < \alpha = 0.01$ for the partial correlation, and of $p\text{-value} > \alpha = 0.01$ for the zero-correlation check (following Sanchez-Romero & Cole (2021)). After the feature selection, we computed FC weights with linear multiple regression and no significance test.

PC algorithm

The Peter-Clark (PC) algorithm (Spirtes et al., 2000; Spirtes & Glymour, 1991) provides a discovery strategy that overcomes the limitations of the three methods described above. It controls for spurious connections created by casual chains, confounders and conditioned-on colliders, even in challenging causal patterns. In addition, after undirected connections are estimated (adjacency discovery phase), the PC algorithm applies a series of orientation rules to

infer, when possible, the causal direction of connections (orientation discovery phase). The outcome of the algorithm is a directed network from which functional causal relationships between brain regions can be read out (**Figure 1F**).

We used an order-independent version of the PC adjacency discovery phase known as PC-stable (Colombo & Maathuis, 2014; termed FAS-stable in Sanchez-Romero et al., 2019), and include pseudocode in **Box 1 (1)**. In short, PC starts with a fully connected undirected network and checks for every pair of brain regions if they are correlated or not. If two regions are *not* correlated, PC removes their adjacency from the network. Next, the algorithm checks for every pair of still connected regions if they are correlated *conditioning on* one other region. (This implies a conditional correlation with a conditioning set of size one. In the pseudocode, the size of the conditioning set **S** is referred to as *depth* (**Box 1 (1).3**).) If two regions are *not* conditionally correlated on one other region, PC removes their connection from the network. For every pair of still connected regions, PC keeps testing correlations conditioning on two other regions (conditioning set of size two), three other regions and so forth, until no more connections can be removed from the network. Note that when PC evaluates conditional correlations, it does it iteratively through all possible combinations of conditioning sets of size one, two, three and so forth, until it finds, if any, a set **S** that makes the regions *not* conditionally correlated (**Box 1 (1).3.b**).

The implementation we used of PC computes the conditional correlation for any two regions X and Y conditioning on a set **S** (**Box 1 (1).3.b.i**), using the inverse of the covariance matrix (precision matrix P) for X, Y and **S**, to obtain the conditional correlation coefficient $r_{XY|\mathbf{S}} = -P_{XY}/sqrt(P_{XX}P_{YY})$, where $sqrt()$ is the square root function and P_{XY} is the entry for X and Y in the precision matrix. To determine statistical significance, first the conditional correlation coefficient $r_{XY|\mathbf{S}}$ is transformed to a Fisher z statistic $z = tanh^{-1}(r_{XY|\mathbf{S}})sqrt(N-|\mathbf{S}|-3)$, where N is the number of datapoints and $|\mathbf{S}|$ is the size of the conditioning set (number of regions conditioned on); then, for a two-sided z-test, we compute the p-value $= 2(1-cdf(abs(z))$, where $abs()$ is the absolute value and $cdf()$ is the cumulative distribution function for a standard normal distribution. For a user-chosen α significance threshold, if p-value $> \alpha$, then we conclude that regions X and Y are *not* correlated conditioning on the set of regions **S**. (In the PC algorithm this result would imply removing the network connection between X and Y.) A significance threshold of $\alpha = 0.01$ was set for all applications of PC.

It is important to note that we implement the PC algorithm with conditional correlations to estimate the required conditional associations, but other approaches can be used, such as conditional mutual information, or other non-linear, non-Gaussian, conditional association measures (Ramsey, 2014; Zhang et al., 2011), depending on the properties of the distributions and functional associations of the data under study

Box 1. PC algorithm pseudocode

(1) Adjacency discovery phase

1. **V** ← set of regions in the input dataset
2. Ŵ ← fully connected undirected network over **V**
3. **for all** *depth* = 0, 1, …, until no more connections can be removed **do**
    a. **for all** *X* in **V do** //guarantees order independence
        i. *a*(*X*) ← set of adjacent regions for *X* in Ŵ
    b. **for all** *X*, *Y*, **S**; *X*, *Y* in **V**, *Y* in *a*(*X*), **S** a subset of *a*(*X*)\{*Y*}, |**S**| = *depth* **do**
        i. **if** *X* is *not* correlated to *Y* conditioning on the set of regions **S then**
            1. remove connection *X* — *Y* in Ŵ
4. **return** Ŵ

(2) Orientation discovery phase

1. **for all** *X* — *Z* — *Y*, and *X* and *Y* not connected in Ŵ **do** //collider orientation
    a. **if** *Z* is *not* in the conditioning set that made *X* and *Y* not correlated **then**
        i. orient *X* → *Z* ← *Y*
2. **for all** *X* → *Z* — *Y*, and *X* and *Y* not directly connected **do** //Meek's rule
    a. orient *Z* → *Y*
3. **return** Ŵ

The second phase of our implementation of the PC algorithm applies two rules to orient, when possible, the adjacencies from the first phase. We include the orientation discovery phase rules in **Box 1 (2)**. The first rule is based on causal principles about conditional independencies implied by collider structures and is part of the original implementation of PC (Spirtes & Glymour, 1991). The collider orientation rule states that if in a network, a region *X* is adjacent to a region *Z*, and *Z* is adjacent to a region *Y*, and *X* and *Y* are *not* adjacent (triple, *X* — *Z* — *Y*), if *Z* is *not* in the conditioning set that made *X* and *Y* not correlated, then necessarily these regions form a collider *X* → *Z* ← *Y* (**Box 1 (2).1**). If the opposite were true, and *Z* were in the conditioning set that made *X* and *Y* not correlated, then we would not be able to orient this triple, because the three possibilities *X* → *Z* → *Y*, *X* ← *Z* ← *Y* or *X* ← *Z* → *Y*, equally imply that *X* is not correlated with *Y* conditioning on *Z*. In this particular case, only collider structures produce unambiguous conditional correlations that can be used to orient adjacencies (see **Figure 1F** for an example).

In practice, when the fMRI time series have a small number of datapoints, some conditional correlation estimates may be inaccurate and we could end up incorrectly orienting colliders. To minimize this risk, Ramsey (2016) implemented the collider orientation rule with the max *p*-value heuristic. The heuristic consists in computing the *p*-values of the correlations of regions *X* and *Y* conditioning on every possible subset of regions adjacent to *X*. Then, choosing the conditioning subset corresponding to the maximum *p*-value, and if region *Z* is *not* in this subset, orienting the triple as a collider *X* → *Z* ← *Y*. By choosing the conditioning subset with the maximum *p*-value we try to guarantee that from all the possible subsets, we select the one that optimally assures that X and Y are conditionally *not* correlated.

Importantly, Ramsey (2016) also showed that with inaccurate conditional correlation estimates due to small number of datapoints, we may end up orienting two conflicting colliders in the network. For example, for two triples we could conclude $X \rightarrow Z \leftarrow Y$ and $Z \rightarrow Y \leftarrow D$, which implies a conflicting orientation for Z and Y. The problem is then which collider orientation should we remove. Ramsey (2016), following the same ideas of the max p-value heuristic, suggests sorting all the previously inferred colliders from high to low according to its p-value—from the max p-value heuristic—and remove a collider orientation if it conflicts with any higher p-value collider. Ramsey (2016) showed in simulations the improvement in orientation accuracy from these two heuristics, so in our implementation of PC we used the collider orientation rule with the max p-value heuristic followed by the collider conflict resolution heuristic.

Meek (1995) introduced a set of orientation rules that in some cases can complement the collider orientations. This second orientation rule (**Box 1 (2).2**) is based on the assumption that the collider orientation rule properly detected all existing colliders in the network, such that no new colliders are allowed. Thus, for $X \rightarrow Z - Y$ we can orient $Z \rightarrow Y$, since the opposite direction $Z \leftarrow Y$ will create a new collider, and that is not allowed. The rest of the Meek's rules leverage the assumption that the underlying causal network does not contain cycles, and thus orient adjacencies avoiding the formation of cycles. Since we know the brain contains feedforward and feedback structures supporting communication between regions, the assumption of no cycles is incorrect in this case. For this reason, we did not implement those orientation rules and prefered to retain undirected connections that may suggest the presence of cycles, than orienting connections based on incorrect assumptions. For mechanistic purposes we consider more problematic a connection oriented in the incorrect causal direction, than no orientation at all.

The output of the PC algorithm is an unweighted directed connectivity network $\hat{W}$ from which connections weights can be estimated. Using $\hat{W}$ as a starting point we derived two different FC approaches. In the first, for each region X, we get $Pa(X)$ the set of causal sources (parents) of X in network $\hat{W}$, and solve the linear regression $X = \beta_X Pa(X)$. The elements of the estimated vector of regression coefficients β_X are considered the weights for the parent connections into X. For example, in $X \rightarrow Z \leftarrow Y$, $Z = \beta_Z Pa(Z) = \beta_{ZX} X + \beta_{ZY} Y$, such that the estimate for β_{ZX} is the weight for the directed connection $X \rightarrow Z$, and equivalently for β_{ZY}. Doing this for every region outputs a FC network, where each directed connection $X \rightarrow Y$ has a causal interpretation in the sense that, keeping all other regions fixed, a change of one unit in X will *cause* a change of β_{YX} in Y (Pearl, 2000; Spirtes et al., 2000; Woodward, 2005). In activity flow models, using a directed FC network implies predicting task-related activity for a held-out region using only its putative causal sources (**Figure 1H**). Hereinafter we refer to this FC method simply as PC algorithm or PC.

The second FC approach is motivated by the hypothesis that the accuracy to predict the activation of a held-out region can increase by using information from both its true direct causal sources and its true direct causal targets (which contain information about the intrinsic processes in the held-out-region that is not provided by the sources) (Aliferis et al., 2010; Fu & Desmarais, 2010). In this approach, for each region X, we get $adj(X)$ the set of adjacent regions

for X in network $\hat{W}$, and solve the linear regression $X = \beta_X adj(X)$. The elements of the estimated vector of regression coefficients β_X are considered the connectivity weights for the adjacencies of X. For example, in $X \rightarrow Y \rightarrow Z$, $Y = \beta_Y adj(Y) = \beta_{YX}X + \beta_{YZ}Z$. Essentially, we are computing the FC weights using every adjacent region, which may include, depending on the inferred causal pattern, only direct causal sources, only direct causal targets or a combination of both—as in the example. These FC weights disregard the orientation of the connections and thus no longer have as straightforward a causal interpretation as in the above PC method. Hereinafter we refer to this FC method as PC-adjacencies or PCadj.

Our implementation of the PC algorithm (with the removal of certain Meek orientation rules as described above) is available at [*project repository will be available upon manuscript acceptance*], and it is a Python wrapper of the PC algorithm from the Java open-source Tetrad software 6.7.1 (available at github.com/cmu-phil/tetrad).

Simulated causal networks and data

As described above, activity flow analysis requires FC estimates from resting-state data and task-evoked activations from task-state data. Our general simulation strategy consists in first creating a synthetic ground-truth causal network (directed graph), simulating dynamics to create a resting-state network and associated dataset, and then simulate a task-state network by introducing small random modifications to the original resting-state coefficients, plus an exogenous task input variable feeding into the network to produce a task-state dataset.

Simulation of resting-state networks and data closely follows Sanchez-Romero & Cole (2021). Networks were based on a directed graphical model that has a preference for common causes and causal chains than for colliders, and includes two-node and three-node cycles. All networks W were simulated with 200 nodes and an average connectivity density of 5% (percentage of connections out of total possible). The connectivity coefficients in W were sampled from a uniform distribution in the interval [0.1, 0.4), and randomly setting 10% of the coefficients to its negative value. To simulate resting-state data we used a causal linear model X = WX+E, where X is a dataset of nodes (nodes × datapoints), W a directed network or matrix of connectivity coefficients (nodes × nodes), with direction encoded from column to row, and E a set of independent noise terms (intrinsic activity) (nodes × datapoints). 1000 datapoints for X were generated by expressing this model as $X = (I-W)^{-1}E$, where I is the identity matrix (nodes × nodes), W is the simulated resting-state network, and pseudo-empirical datapoints for E (Sanchez-Romero & Cole, 2021) were produced by randomizing preprocessed fMRI resting-state data across datapoints, regions and participants, from the Human Connectome Project (HCP). Using pseudo-empirical terms E allowed us to simulate resting-state data X that better capture some of the distributional properties of the empirical fMRI.

To create task-state networks W_T, we took the previously simulated resting-state networks W and defined with equal probability each task-state coefficient as: (a) one standard deviation above the corresponding resting-state coefficient, or (b) one standard deviation below, or (c) equal to the resting-state coefficient. This ensured that the connectivity coefficients for rest and task-state were not exactly the same. To generate task-state data we first introduced an

exogenous task input variable T with pseudo-empirical datapoints (1 × datapoints). Then, we defined a task connectivity vector C (nodes × 1), that specifies the network nodes *directly* affected by the task variable T. We randomly chose 10% of nodes to be directly affected by T, and sampled task connectivity coefficients from a uniform distribution in the interval [0.1, 0.4). (Note that other nodes can also be affected by T but through indirect causal paths.) Expanding the linear model to include the task-related elements, we now have $X_T = W_T X_T + CT + E_T$. We generated 1000 datapoints for task-state dataset X_T by expressing the model as $X_T = (I-W_T)^{-1}(CT+E_T)$. As with the resting-state data we used pseudo-empirical datapoints for E_T.

Finally, simulated task-evoked activations were computed individually for each task-state network node X_T by a linear regression of the form $X_T = aT$, where the estimated coefficient a represents the task-evoked activation, and reflects direct and indirect effects of the task variable T on the node X_T.

200 instantiations of the simulated models were generated to compare (1) the accuracy of the different FC methods to recover resting-state networks, and (2) the prediction accuracy of the activity flow models parameterized with these networks. Analyses were run in the Rutgers-Newark high-performance computing cluster AmarelN (oarc.rutgers.edu/resources/amarel), using one node, 2 cores, and 64G RAM.

Empirical fMRI data

We used open access fMRI resting and task-state data from a subset of 176 participants from the minimally-preprocessed HCP 1200 release (Glasser et al., 2013; Ugurbil et al., 2013; Van Essen et al., 2013). All subjects gave signed informed consent in accordance with the protocol approved by the Washington University institutional review board. We abide by the HCP open access use terms and the Rutgers University institutional review board approved use of these data. These participants were selected by passing the following exclusion criteria described in Ito, Brincat, Siegel, et al. (2020): anatomical anomalies found in T1w or T2w scans; segmentation or surface errors as output from the HCP structural pipeline; data collected during periods of head coil problems; data in which some of the FIX-ICA components were manually reclassified; participants that had any fMRI run in which more than 50% of TRs had greater than 0.25mm motion framewise displacement; removal according to family relations (only unrelated participants were selected, and those with no genotype testing were excluded). We include here a brief description of the data fMRI collection parameters: whole-brain echo-planar functional imaging acquisitions were acquired with a 32 channel head coil on a modified 3T Siemens Skyra MRI with TR = 720 ms, TE = 33.1 ms, flip angle = 52°, BW = 2290 Hz/Px, in-plane FOV = 208 × 180 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. For our analysis, we only used one 14.4 minutes run of resting-state data (1200 datapoints), and two 30 minutes consecutive runs (60 min total) of task-state data (7 tasks with 24 conditions). Further task and resting-state data acquisition details can be found elsewhere (Barch et al., 2013; Smith et al., 2013).

In brief, the seven tasks consisted of an emotion cognition task (valence judgment, 2 conditions); gambling reward task (card guessing, 2 conditions); language processing task (2

conditions); motor task (tongue, finger, toe, 6 conditions); relational reasoning task (2 conditions); social interaction cognition task (2 conditions); and working memory task (0-back, 2-back, 8 conditions). Details about these task paradigms can be found in Barch et al. (2013).

The minimally-preprocessed HCP surface data (Glasser et al., 2013) were first parcellated into 360 cortical regions using the Glasser et al. (2016) atlas. Then, we applied the preprocessing steps detailed in Ito et al. (2020). Briefly, they include removing the first five datapoints of each run, demeaning and detrending the time series, and nuisance regression—based on Ciric et al. (2017)—with 64 parameters to control for the effects of motion and physiological artifacts, and their derivatives and quadratics. Global signal regression was not applied since its physiological basis and effects on functional connectivity inferences are still not fully understood (Aquino et al., 2020; Colenbier et al., 2020; Li et al., 2019; T. T. Liu et al., 2017; Murphy & Fox, 2017).

Task-evoked activations for each of the 360 regions and 24 conditions were estimated using a standard general linear model at the region level. The SPM software canonical hemodynamic response function (fil.ion.ucl.ac.uk/spm) was used for general linear model estimation , given that all tasks involved block designs (Cole et al., 2021).

Resting and task-state empirical fMRI data were used to compare the accuracy of activity flow predictions under the five different FC methods tested here. Analyses were also run in the AmarelN cluster mentioned above.

Data and code to reproduce our synthetic and empirical analyses are available at [*project repository will be available upon manuscript acceptance*].

Results

Network recovery on simulated fMRI data

We began by simulating ground-truth functional causal networks to determine the validity of FC methods, each of which we hypothesized to be at different levels of causal validity (**Figure 1A**). A series of 200 random networks, each with 200 nodes, were simulated from a graphical causal model with more common causes and causal chains than colliders, and two-node and three-node cycles. Positive and negative connectivity weights were sampled from a uniform distribution. For each causal network, we simulated fMRI time series with 1000 datapoints using a linear model and randomized empirical resting-state fMRI data. The validity of the FC methods was assessed in terms of how good they recovered ground-truth resting-state networks. We used precision and recall as measures of recovery accuracy.

We first report in **Figures 2A-C**, precision and recall for the recovery of the true network adjacency pattern in simulated data, for each of the FC methods tested. Precision is defined as the number of true-positive adjacencies (*tp*) divided by the sum of the number of true-positive and false-positive adjacencies (*fp*) ($precision = tp/(tp+fp)$). Precision values range from 0 to 1, and quantify the ability of each FC method to avert false-positive (spurious) connections. A

precision of 1 indicates that the method did not output any false positives, and a precision of 0 that it only output false-positive connections. Recall is defined as the number of true-positive adjacencies divided by the sum of the number of true-positive and false-negative adjacencies (*fn*) (*recall* = *tp*/(*tp*+*fn*)). Recall values also range from 0 to 1, and reflect the ability of each FC method to recover true connections. A recall of 1 indicates that the method inferred all the true connections, and a recall of 0 that it did not recover any of the true connections. Together, precision and recall yield a complementary view of each method's capacity to recover the true network while avoiding spurious connections. Results are reported in boxplots indicating median, and lower and upper quartiles for 200 simulated resting-state networks.

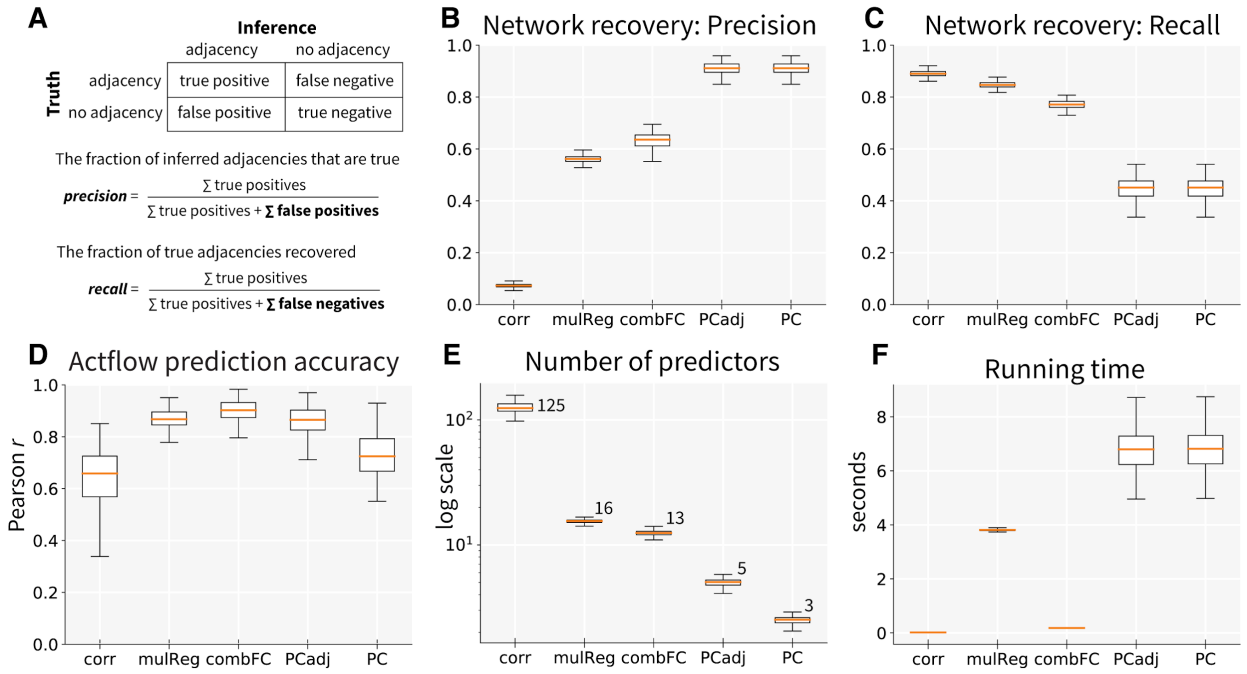

Figure 2 — Recovery of functional connectivity networks and accuracy of activity flow prediction of task-evoked activations for simulated fMRI data. Boxplots show median, and lower and upper quartiles for 200 simulations. Correlation (corr), multiple regression (mulReg), combinedFC (combFC), PC-adjacencies (PCadj) and PC algorithm (PC). (**A**) Precision and recall formulas to measure goodness of network adjacency recovery. (**B**) Precision. (**C**) Recall. (**D**) Accuracy of actflow prediction of task-evoked activations, measured with Pearson correlation coefficient *r* and averaged across 200 regions. (**E**) Number of predictor regions averaged across 200 regions, plotted on a logarithmic scale for visualization, with actual median values next to each boxplot. For reference, the median in-degree (number of direct causal sources) in the true networks is 6.62. (**F**) Running time in seconds.

Spurious causal inferences (false positives) are generally considered to be more problematic than false negatives in science. This strategy reduces the chance that a false result will be considered a scientific fact. With the primary goal of avoiding spurious causal inferences, we sought to maximize precision first (reducing false positives), followed by maximizing recall

(reducing false negatives). Thus, the method with the highest precision would be preferred even with a moderately lower recall.

Considering the theoretical properties of each FC method to avoid spurious connections from confounders, chains and conditioned-on colliders, we expect the PC algorithm to achieve the highest precision (less false-positive connections) for network recovery, followed by combinedFC, multiple regression and correlation, in that order. Regarding recall, it is not clear which of the FC methods is expected to recover the highest number of true connections.

As expected, **Figure 2B** shows that PCadj and PC dominated over the rest of the methods with a median precision of 0.91, indicating that 91% of its inferred adjacencies are true positives (true connections) and 9% are false positives (spurious connections). Note that PCadj and PC have the exact same precision since PCadj is PC without the orientation information, and here precision is calculated only for adjacencies. CombinedFC and multiple regression were next, with precisions of 0.64 and 0.56 respectively. The lowest scoring method was correlation, with a precision of 0.07, implying that 7% of its inferred adjacencies are true positives and 93% are false positives.

Figure 2C shows that the results for recall go in the opposite direction. Correlation had the highest median recall, 0.89, indicating that 89% of the total true connections were recovered. Multiple regression and combinedFC follow closely with 85% and 77% respectively, while PCadj and PC had the lowest recall, with 45% of the true connections recovered.
From these methods only the PC algorithm has inference rules to orient connections (**Box 1 (2)**). To measure the orientation accuracy of PC, meaning if the causal direction of an adjacency was correctly inferred, we computed the proportion of correct orientations out of the total number of correctly inferred adjacencies (true-positive adjacencies). Notably, PC showed a median orientation accuracy of 0.83 across the 200 simulations, implying that 83% of the true-positive adjacencies were oriented in its correct causal direction.

These first results confirm that FC methods grounded in stronger causal principles, such as combinedFC or PC, can recover with higher precision (lower number of spurious connections inferred) simulated resting-state fMRI networks. This suggests that more causally-valid FC methods will result in activity flow models with better prediction accuracy, since the prediction of the task-evoked activations will not contain effects from spurious or indirect pathways.

Actflow prediction accuracy on fMRI synthetic networks

We extended our ground-truth models to determine if more causally valid FC methods, by better controlling the effect of spurious connections, lead to more accurate predictions of task-evoked activity. Task-state networks were simulated by taking the previous 200 resting-state simulated networks and applying minor variations to the original connectivity weights. For each iteration, the task was modeled as an exogenous source node, randomly connected to a number of network nodes. Task fMRI time series with 1000 datapoints were simulated for each network. We regressed each task-state network node on the exogenous task and considered the regression coefficients as the to-be-predicted task-evoked activations.

In **Figure 2D** we report the prediction accuracy of the actflow models for each of the FC methods tested. Following Cole et al. (2016), we measured prediction accuracy with the Pearson correlation coefficient *r* between the vector of predicted activations for the 200 simulated regions and the vector of actual activations. Measured in this way, the prediction accuracy *r* summarizes the actflow model performance across the whole network. Boxplots show median, and lower and upper quartiles for 200 simulations.

Actflow predictions are a function of the estimated FC network and the actual task-evoked activations, therefore we expect that a better recovery of the true network will lead to a higher actflow prediction accuracy. Previous fMRI empirical results (Cole et al., 2016) showed that functional networks estimated with multiple regression produced more accurate actflow predictions compared to networks estimated with correlation. In addition, our simulation results (**Figures 2A** and **2B**) show that multiple regression-based networks have a higher number of false negatives (lower recall) but a lower number of false positives (higher precision) than correlation-based networks. Together, these observations suggest that actflow prediction accuracy may be more improved by reducing false-positive functional connections than by reducing false-negative connections. Thus, we expect that PCadj and PC—the methods with the best network recovery precision—will have the higher prediction accuracy, followed by combinedFC, multiple regression and correlation, in that order.

CombinedFC had the best median prediction accuracy of all the methods (r = 0.90), probably due to a good balance of false positives and false negatives. It was followed closely by PCadj (r = 0.87) and multiple regression (r = 0.87). PC had a lower prediction accuracy compared to these methods (r = 0.72). Note that when the actflow model is parameterized with the PC-based oriented network, it only uses the estimated direct sources to predict the activity of each held-out region. In contrast, when actflow uses unoriented networks, derived from multiple regression, PCadj or combinedFC for example, it considers all the adjacent regions of a held-out region to predict its activity. In causal terms, this implies that with unoriented networks, the held-out region activity prediction leverages information from both direct sources (forward causation) and direct targets (backward causation), achieving a better prediction than when only source information is used. It is for this reason that the PC prediction accuracy was lower than PCadj, multiple regression and combinedFC prediction accuracies. Critically, the higher network recovery precision for the PC-based models (**Figure 2B**) means that (despite having lower prediction accuracy) they are more causally valid, increasing mechanistic interpretability of the actflow-generated activity predictions.

As expected, actflow models based on correlation networks showed the lowest accuracy of all (r = 0.66). Despite the low number of false negatives (high recall), these models have a high number of false-positive connections (low precision), which create spurious pathways through which task activity is incorrectly accounted for.

Note that a vector of predictions and a vector of actual activations can have a high Pearson *r*, even if their values are in a completely different scale (e.g., all values multiplied by 2). If we are interested in assessing the deviation from the actual values, we can use the coefficient of

determination $R^2 = 1-(\Sigma_i(A_i-\hat{A}_i)^2/\Sigma_i(A_i-\bar{A})^2)$, where A_i are the actual activations, $\hat{A}_i$ the predicted activations, and $\bar{A}$ the mean of the actual activations. R^2 measures the proportion of the variance of the actual activations that is explained by the prediction model. It ranges from 1 (perfect prediction) to minus infinity (prediction deviations can be arbitrarily large), with a value of zero when the predictions are equal to the mean of the actual activations ($\hat{A}_i = \bar{A}$). For our simulations, the median actflow prediction R^2 for the FC methods followed the same ordering as the Pearson r: combinedFC (R^2 = 0.80), PCadj (R^2 = 0.74), multiple regression (R^2 = 0.68), PC (R^2 = 0.52) and correlation (R^2 = −164.68). The high negative R^2 of correlation-based models indicates that predicted values strongly deviate from actual activation values, confirming the detrimental effect of spurious network pathways on the actflow prediction model.

The complexity of actflow models can be evaluated with the number of regions used to predict the task-evoked activation of held-out regions. **Figure 2E** shows the number of predictors for each held-out region, averaged across the 200 regions. Notably, actflow models based on PC causal networks were highly accurate (**Figure 2D**), despite having the lowest model complexity (median of 3 predictors, since only causal sources were used to predict activations). These results evidence that PC can successfully recover functional causal connections that have high predictive power. This is further confirmed with the results of PCadj-based actflow models, which also had a relatively low number of predictors (median of 5), and an accuracy as high as combinedFC and multiple regression, both with an order of magnitude more predictors (median of 13 and 16 respectively). Correlation-based actflow models reported the highest complexity (median of 125 predictors) and the lowest accuracy to predict task-evoked activations.

Finally, **Figure 2F** shows that all FC methods have efficient running times which do not exceed the tens of seconds. The PC algorithm had the longest median running time (7 sec), which is very efficient considering the large number of conditional associations it has to compute due to the number of regions and the complexity of the connectivity patterns in the true networks. Surprisingly, multiple regression had a relatively long median running time (4 sec). Analysis of our code revealed that this unexpected running time was caused by the significance test for the multiple regression coefficients. Not computing the significance test considerably reduces the running time. Running times depend on the hardware used for the analysis, but we expect that the reported ordering of the methods will replicate in any machine.

Our results on simulated fMRI data show that causally-valid FC methods, such as PC and combinedFC, can be used to build activity flow models with high prediction accuracy, low complexity (number of predictors) and efficient running times, thanks to their use of causal principles to control for the effects of confounders, causal chains and conditioned-on colliders, and if orientation rules are provided, such as with PC, these methods can provide mechanistic interpretations of the predicted task-evoked activations.

Actflow prediction accuracy on fMRI empirical networks

Our theoretical considerations about the causal validity of the FC methods (**Figure 1A**) and the comparative performance observed above in simulations, prompted us to hypothesize that this performance will translate, up to a degree, to an empirical domain. Here, we used fMRI data to

assess the performance of actflow predictive models under a complex empirical setup—for which the ground-truth networks are not known—comprising 360 cortical regions, 24 task conditions across various cognitive domains and 176 different participants. We compared the average prediction accuracy of the different FC methods across all conditions.

Formally, prediction accuracy was measured as the Pearson correlation coefficient r between the vector of predicted activations for all 360 cortical regions of Glasser et al. (2016) and the vector of actual activations, averaged across the 24 HCP task conditions. Measured in this way, the prediction accuracy r summarizes the actflow model performance across the whole-brain and the full set of task conditions. **Figure 3A** reports actflow prediction accuracy boxplots with median, and lower and upper quartiles across the 176 participants, for each connectivity method. PCadj attained the highest median prediction accuracy (r = 0.82), followed by combinedFC (r = 0.77) and PC (r = 0.74). Multiple regression (r = 0.60) and correlation (r = 0.57) showed lower accuracies. As in the simulation results, we include median coefficient of determination R^2 as a complementary measure of prediction accuracy: PCadj (R^2 = 0.67), combinedFC (R^2 = 0.59), PC (R^2 = 0.53), multiple regression (R^2 = 0.33) and correlation (R^2 = −710). As remarked in the simulation results, the high negative R^2 of the correlation-based models reflects large differences between the predicted and the actual activation values, which are likely the result of spurious network pathways. As we hypothesized, these empirical results are consistent with the simulations, in the sense that actflow models parameterized with more causally valid FC methods, such as PC and combinedFC, can better predict task-evoked activity.

The number of predictors in the actflow models (**Figure 3B**) followed the same order observed in simulations. PC had the lowest model complexity (median of 3 predictor source regions), followed by PCadj (median of 6), combinedFC (median of 8), multiple regression (median of 10) and correlation (median of 253). These results also reflect the different connectivity densities of the estimated FC networks. Correlation produced highly dense networks (**Figure 3D**), resulting in actflow models with high complexity (large number of predictors) but low prediction accuracy (**Figure 3A**), whereas PC inferred sparser networks (**Figures 3G-H**), that produced actflow models with lower complexity and high prediction accuracy.

Lastly, **Figure 3C** reports running times for the inference of FC networks from empirical fMRI time series. We observed efficient running times not exceeding tens of seconds. For these data, multiple regression showed the longest median running time (44 seconds), followed by PC (33 seconds). As mentioned above, the relatively long running time of multiple regression comes mainly from the computation of the significance test for the regression coefficients.

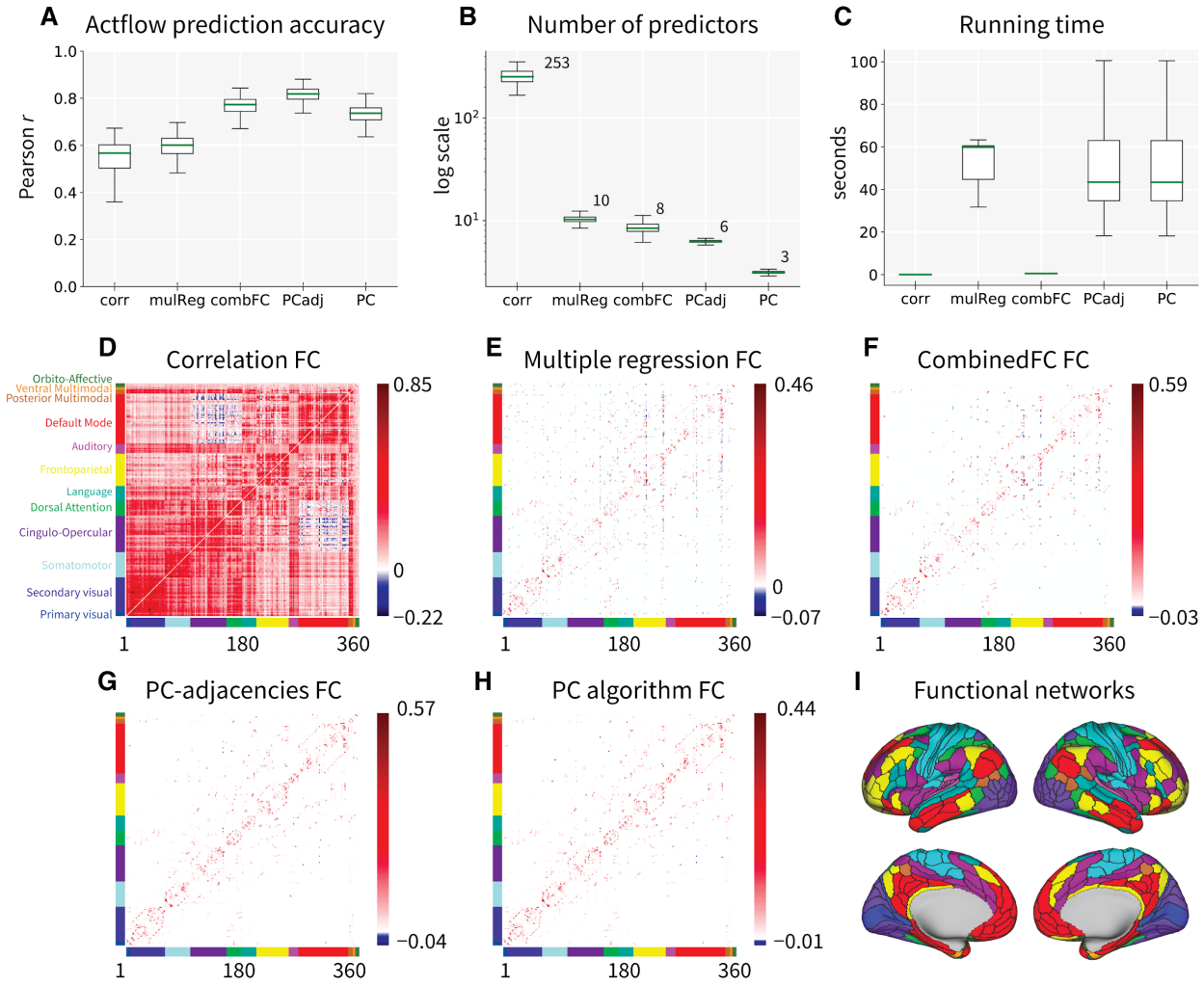

Figure 3 — Accuracy of activity flow prediction of task-evoked activations and functional connectivity networks for fMRI empirical data. Correlation (corr), multiple regression (mulReg), combinedFC (combFC), PC-adjacencies (PCadj) and PC algorithm (PC). Boxplots show median, and lower and upper quartiles for the 176 participants. (**A**) Prediction accuracy for activity flow models parameterized with each of the FC methods tested. Accuracy was measured as the Pearson correlation coefficient *r* between the vector of predicted activations for all 360 cortical regions of Glasser et al. (2016) and the vector of actual activations, averaged across the 24 HCP task conditions. (**B**) Number of predictor regions averaged across 360 regions, plotted on a logarithmic scale for better visualization, with actual values next to each boxplot. (**C**) Running time in seconds for network estimation. (**D**) Correlation-based functional connectivity (FC) network, averaged across 176 participants. The 360 cortical regions were organized in 12 functional networks from Ji et al. (2019). (**E**) FC average network from multiple regression. (**F**) FC average network from combinedFC. (**G**) FC average network from PC-adjacencies. (**H**) FC average network from PC algorithm. (**I**) Cortical surface map of the 12 functional networks partition used in panels D-H (available at github.com/ColeLab/ColeAnticevicNetPartition). For visualization, in all FC networks (panels D-H), values between -/+ 0.005 were set to zero.

Together, these results confirm empirically that causally-informed methods can estimate undirected and directed FC networks useful to build activity flow models with low complexity

and high prediction accuracy. Actflow models using undirected FC networks can provide excellent predictions (e.g., combinedFC and PCadj), but cannot tell if the task activity is flowing into, from, or into and from (as in a feedback) the held-out region. In contrast, PC-based directed networks, despite some reduction in the prediction accuracy, provide mechanistic models in which the flow of task activity can be traced from source regions directly into held-out regions.

Actflow predictions across brain regions and task conditions

Our results so far have confirmed in simulations and empirically (**Figure 3A**, summary across brain regions, task conditions and participants) the benefits in prediction and mechanistic interpretation from more causally-grounded FC methods. Notably, we showed that directed FC networks from the PC algorithm can be used to build directed actflow models (**Figure 1H**) with high prediction accuracy, low complexity and a tractable mechanistic interpretation. We now extend these results by comparing and visualizing at more detail, the performance of PC-based directed actflow models against actflow models parameterized with the field-standard Pearson correlation FC. This analysis is essentially focused on measuring the accuracy in predicting a whole-brain pattern of regional activations for each task condition (node-wise accuracy) (Cole et al., 2021).

The actflow accuracies reported in **Figure 3A** summarize on the Pearson *r*, general information from all 360 cortical regions and 24 task conditions. Here, we unpacked this information by showing in matrix form the actual task activations for all regions (rows) and task conditions (columns), median across the 176 participants (**Figure 4B**), flanked by the predicted activations of the field-standard Pearson correlation-based models (**Figure 4A**) and by the predictions of the PC-based mechanistic actflow models (**Figure 4C**). These matrices show with greater detail how PC-based predictions across regions and all conditions had a more similar pattern to the actual activations, relative to correlation-based predictions ($r = 0.74 > r = 0.57$, median across participants). Importantly, we also observed that the PC-based predicted values were in approximately the same range as the actual activations (compare colorbars in **Figures 4B** and **4C**), while correlation-based predicted activations were often hundreds of times off of the actual values (see **Figure 4A** colorbar). As mentioned above, these deviations can be quantitatively assessed with the coefficient of determination, $R^2 = 0.53$ vs. $R^2 = -710$, for PC-based models and correlation-based models correspondingly. The high negative R^2 reflects the observed strong deviations in the correlation-based actflow predicted values.

For each of the 24 task conditions taken individually, we confirmed that PC-based actflow models attained a significantly higher node-wise prediction accuracy than correlation-based models (for both Pearson *r* and R^2, $p<0.01$ corrected for multiple comparisons with the nonparametric test of Nichols & Holmes (2002), for 176 participants).

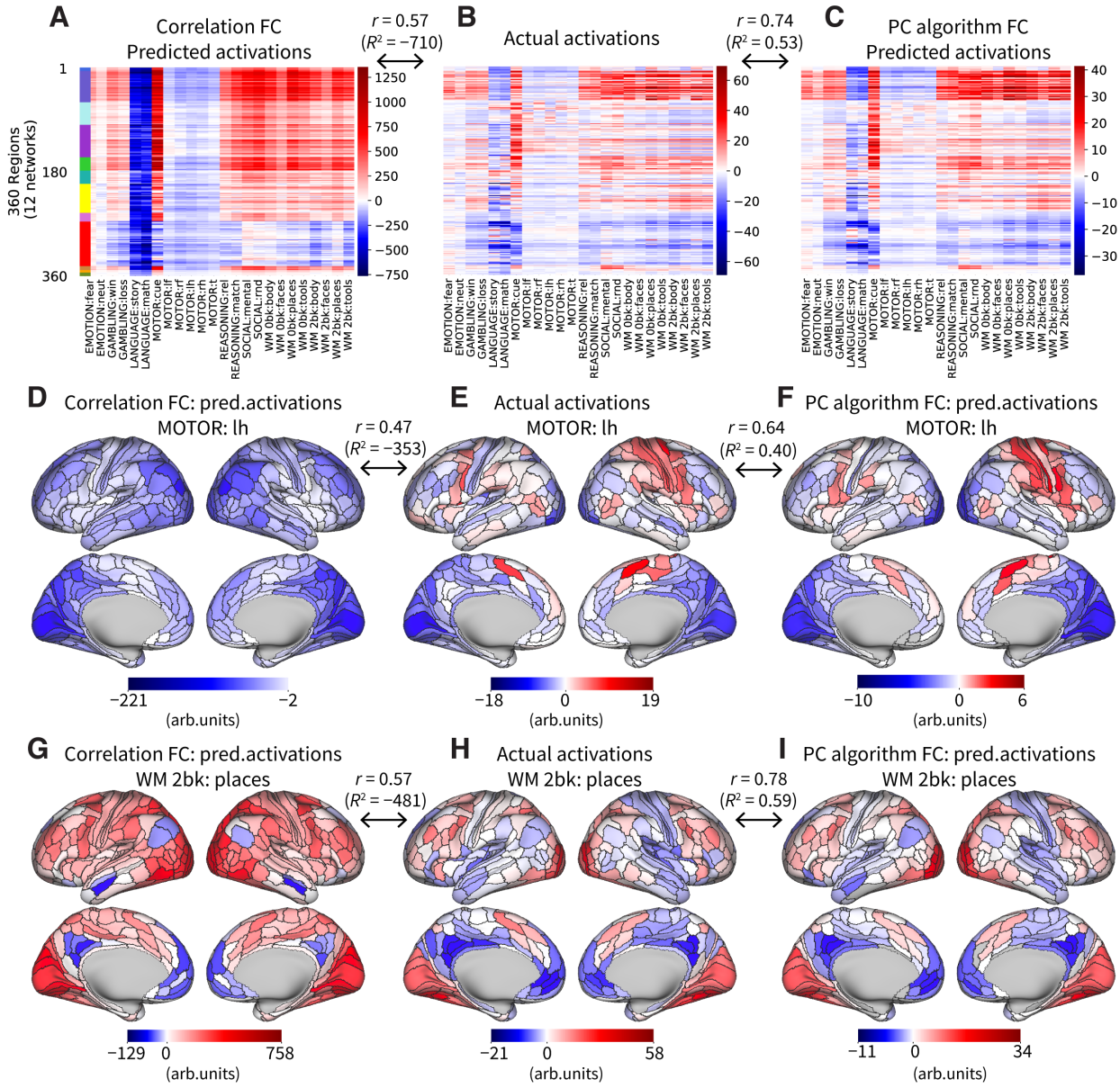

Figure 4 — Activity flow prediction of task-evoked activations for fMRI empirical data, across brain cortical regions and task conditions. (**A**) Actflow predicted activations based on correlation FC, for 360 regions organized in 12 networks (rows) (Glasser et al., 2016; Ji et al., 2019), and 24 HCP task conditions (columns), median across 176 participants. The median r = 0.57 (R^2 = −710) results from first computing the Pearson r (coefficient of determination R^2) prediction accuracy node-wise (360 regions) for each of the 24 conditions, then averaging across conditions and finally obtaining the median across participants. (**B**) Actual task-evoked activations. (**C**) Actflow predicted activations based on PC algorithm FC. Median prediction accuracy of r = 0.74 (R^2 = 0.53). (**D**) Correlation-based actflow predictions for the motor task: left hand condition, projected into a brain surface map with 360 Glasser regions, median across 176 participants. The prediction accuracy of r = 0.47 (R^2 = −353) is the Pearson r (coefficient of determination R^2) between the vector of node-wise (360 regions) predicted activations for this condition and the vector of actual activations, median across participants. (**E**) Actual activations for the motor task: left hand condition. (**F**) PC-based actflow predictions for the motor task: left hand condition. Median prediction accuracy of r = 0.64 (R^2 = 0.40). (**G**) Correlation-based actflow predictions for the working memory task: 2-back: places condition, median across 176 participants. Median

prediction accuracy of $r = 0.57$ ($R^2 = -481$). (**H**) Actual activations for the memory task: 2-back: places condition. (**I**) PC-based actflow predictions for the memory task: 2-back: places condition. Median prediction accuracy of $r = 0.78$ ($R^2 = 0.59$).

To further illustrate this result, we performed a more targeted analysis that focuses on one task condition (one column in the **Figures 4A-C** matrices) and examines the pattern of actflow predicted activations across the 360 brain regions (node-wise accuracy). We first show this analysis for the motor task: left hand condition. The PC-based actflow models (**Figure 4F**) recovered the whole-brain actual activation pattern (**Figure 4E**) better than the correlation-based models (**Figure 4D**) ($r = 0.64 > r = 0.47$; and $R^2 = 0.40$ vs. $R^2 = -353$, median across participants). More importantly, the PC-based models correctly recovered the functionally-relevant pattern of positive activations in the right hemisphere somatomotor network, known to be engaged during task-induced left hand movements. In contrast, correlation-based models predicted negative and inflated activation values across the entire cortex.

We repeated the node-wise accuracy analysis, this time for the working memory task: 2-back: places condition. In this case, we also confirmed the superior performance of PC-based actflow models (**Figure 4I**) compared to correlation-based models (**Figure 4G**), both in the prediction of the whole-brain activation pattern and in the range of predicted values ($r = 0.78 > r = 0.57$; and $R^2 = 0.59$ vs. $R^2 = -481$, median across participants). Note how the predictions of correlation-based models are massively biased towards positive and negative values (**Figure 4G**, colorbar), reflecting again the presence of inferred spurious pathways through which task activity is incorrectly summed to the actflow computation.

These results confirm that mechanistic actflow models based on causally-valid directed FC methods (such as PC, that controls for indirect and spurious pathways, and provide causal source information), better predict whole-brain activation patterns and actual values, for each of the 24 task conditions tested here. In contrast, correlation-based FC produced densely connected actflow models in which every region's connectivity profile probably conflated direct, indirect and spurious pathways (see **Figure 3D** FC matrix) that incorrectly biased the actflow predicted activations.

Focal activity flow prediction across task conditions

The previous analysis compared the accuracy of PC- and correlation-based actflow models in predicting whole-brain activation patterns for individual task conditions (node-wise accuracy, e.g., **Figure 4F**). Here, in contrast, we want to measure the accuracy of PC-based directed actflow models and correlation-based models for predicting activations across the 24 task conditions for each individual brain region (condition-wise accuracy) (Cole et al., 2021). This analysis allows us to highlight brain areas for which the two methods differ significantly.

We consider the matrices in **Figures 4A-C**, choosing one row (region) and computing the Pearson r value between the vector of 24 column (conditions) actual activations and the vector

of 24 column actflow-predicted activations. The prediction accuracy *r* value for each region is then projected to a brain map. **Figures 5A-B** show the condition-wise accuracy of correlation-based and PC-based actflow models for each region (median across 176 participants).

To highlight differences in prediction accuracy across the methods, **Figure 5C** shows the PC-based condition-wise accuracy minus the correlation-based accuracy for each brain region. For 82% of the 360 regions PC-based actflow models attained a significantly higher condition-wise prediction accuracy than correlation-based models (Pearson *r*, $p<0.01$ corrected for multiple comparisons with the nonparametric test of Nichols & Holmes (2002), for 176 participants. 99% of the 360 regions for R^2). The brain map of **Figure 5C** shows that accuracy differences are relatively larger in the language and somatomotor networks. This condition-wise result is consistent and complementary to the previous node-wise accuracy analysis of the motor and working memory tasks (**Figures 4D-I**).

These results extend our previous observations by confirming that PC-based directed actflow models (with lower complexity and valid mechanistic interpretation) can better predict, for a large majority of cortical regions, the task-evoked activations across a diverse set of cognitive conditions.

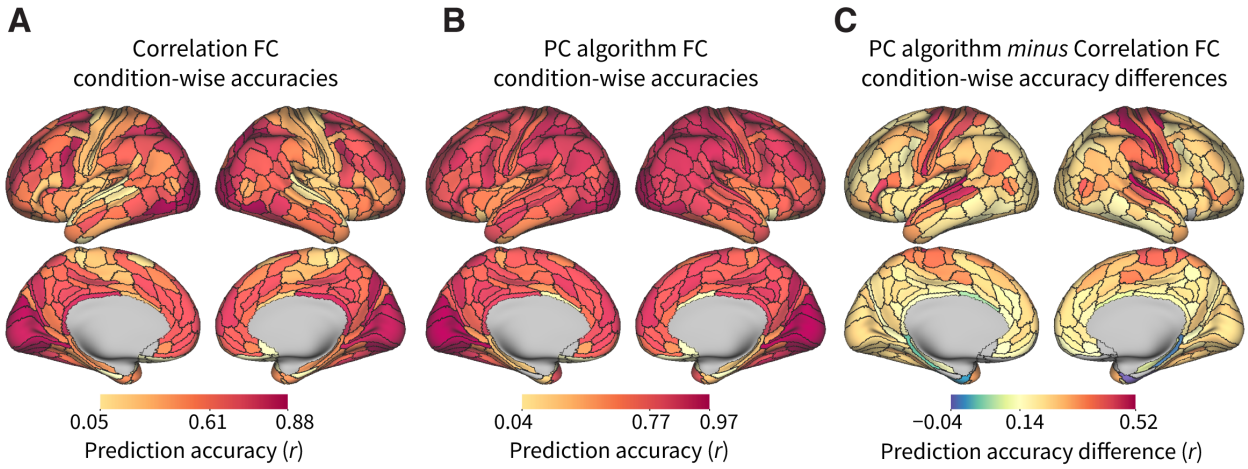

Figure 5 — Condition-wise accuracy of activity flow predictions for 360 cortical regions. Condition-wise actflow accuracy is the prediction accuracy for each region across the 24 task conditions of the HCP data. Brain surface maps show median across 176 participants. Colorbars indicate minimum, mean and maximum prediction accuracy (*r*) across the 360 Glasser regions. (**A**) Actflow condition-wise accuracies using the field-standard Pearson correlation FC. (**B**) PC algorithm-based actflow condition-wise accuracies. (**C**) PC algorithm-based minus correlation-based actflow condition-wise accuracies. Positive values indicate that PC algorithm-based actflow models predict better than correlation-based models, negative values indicate the opposite.

Directed activity flow models improve mechanistic insight into task condition selectivity

Results thus far demonstrate that parameterizing activity flow models with causally-principled FC methods improves our ability to predict task-evoked activations. This was shown via summaries of results across all regions and task conditions. Here, we focus on one particular region engaged in a single cognitive task manipulation to illustrate how causal actflow models can be used to provide more precise mechanistic insight into the emergence of task-evoked activations of interest. Specifically, we applied an actflow model to understand how the flow of activity from distal brain regions may give rise to an established cognitive activation effect (2-back vs. 0-back working memory task conditions) in one region of the right dorsolateral prefrontal cortex (DLPFC). We focused on a cognitive contrast given the enhanced causal experimental control inherent in such a contrast (controlling for, e.g., stimulus perception, task timing, general task engagement). Results are compared for correlation-based and PC-based actflow models.

First, activations for the 4 conditions for 2-back (body, face, place and tool stimuli) (see **Figures 4A-C**) were averaged to form one 2-back activation for each of the 360 Glasser regions. The same procedure was applied for the 0-back conditions. Then, activation differences or contrasts (2-back minus 0-back) were computed, and regions with higher 2-back than 0-back activation were localized (**Figure 6A**, red regions).

Brodmann's Area 8 in the posterior section of the DLPFC has been extensively reported to have an important role in the maintenance of stimuli information during working memory tasks (Carlson et al., 1998; Constantinidis et al., 2001; Courtney et al., 1998; Petrides, 2000; Rowe et al., 2000; Rowe & Passingham, 2001; Wager & Smith, 2003). In the Glasser et al. (2016) parcellation, DLPFC Area 8 is subdivided into areas 8C, 8Av, 8Ad and 8BL, from which the right hemisphere Area 8C showed the highest positive activation in the 2-back vs. 0-back contrast—both in the analysis conducted here and the one in Glasser et al. (2016). The right DLPFC Area 8C therefore has a prominent and established role in working memory function, and this motivated us to choose it as our to-be-predicted target (**Figure 6C**, green region).

We built an actflow model for right Area 8C using the actual activation contrasts for the rest of the regions (**Figure 6A**) and the estimated resting-state functional connections (**Figures 6B-C**). Multiplying each region's actual activation contrast by its corresponding connection to the to-be-predicted target, we obtained a brain map of activity flow contrast estimates (**Figures 6D-E**). Positive values indicate incoming contributions that would increase activity in the DLPFC Area 8C, while negative values indicate incoming contributions that would decrease activity in this region. Finally, we summed the activity flow estimates to predict the task-evoked activation contrast for the chosen DLPFC region.

As a consequence of Pearson correlation's inability to control for spurious effects of confounders and causal chains, the correlation-based actflow models used a considerably high number of

regions in the left and right hemispheres (**Figure 6B**) to predict the activation contrast of the right DLPFC Area 8C. These densely connected models predicted an activation contrast one order of magnitude larger than the actual contrast (52.22 vs. 8.3, average across 176 participants) (**Figure 6D**).

In addition to its inaccurate prediction, the correlation-based actflow model is poorly informative in a mechanistic sense, since we cannot derive reliable causal hypotheses about the effects of possible interventions. For example, given the lack of oriented connections, and the presence of uncontrolled confounders and chains, intervening on a highly correlated region does not guarantee that the target region will be affected with the assumed strength and sign, or even affected at all.

In contrast, the PC algorithm inferred a sparser and more causally-valid directed connectivity pattern for the DLPFC Area 8C, with sources in parietal, frontal and temporal lobes, principally on the right hemisphere (**Figure 6C** and **Table 1**). This sparser causal pattern resulted in a causal actflow prediction closer to the actual contrast value (5.45 vs. 8.3, average across 176 participants) (**Figure 6E**).

Most importantly, theory and simulation results both suggest that the activity flows computed using PC are more causally valid, such that they can be used to infer the likely direct causes of activity in a given neural population. In this case, it is likely that a specific set of cortical regions (visualized in **Figure 6E**) contribute to the emergence of working memory effects in Area 8C of the right DLPFC. We detected 95 regions that on average across participants, have a non-zero contribution to the PC-causal DLPFC prediction. After ordering them according to the strength of their contribution, we found (after increasing the number of removed regions one-at-a-time) that by collectively removing the top 21 regions from the actflow computation (simulated lesioning) (**Table 1**), the predicted working memory contrast becomes statistically non-significant (average across participants 0.16, p-value>0.10 for a two-sided t-test), which strongly supports the relevance of these functional interactions. (The full list of 95 contributing regions is included as Supplementary Table 1.)

This analysis goes beyond previous studies that have determined that DLPFC is especially active during n-back tasks (Evangelista et al., 2021; Kumar et al., 2017; Sherwood et al., 2016; Woodcock et al., 2018) and studies that have characterized DLPFC connectivity (Cole et al., 2012; Panikratova et al., 2020; Reineberg & Banich, 2016), revealing likely direct causal mechanisms contributing to DLPFC's involvement in n-back task cognitive processes. As mentioned before, this single-causal-step activity flow model represents a starting point for a more comprehensive multiple-causal-step neurocognitive explanation of working memory. This explanation would characterize the causal chain from stimulus (i.e., n-back task) to network-based cognitive phenomenon (i.e., neural activity changes from differences in working memory load) to behavior (i.e., motor response).

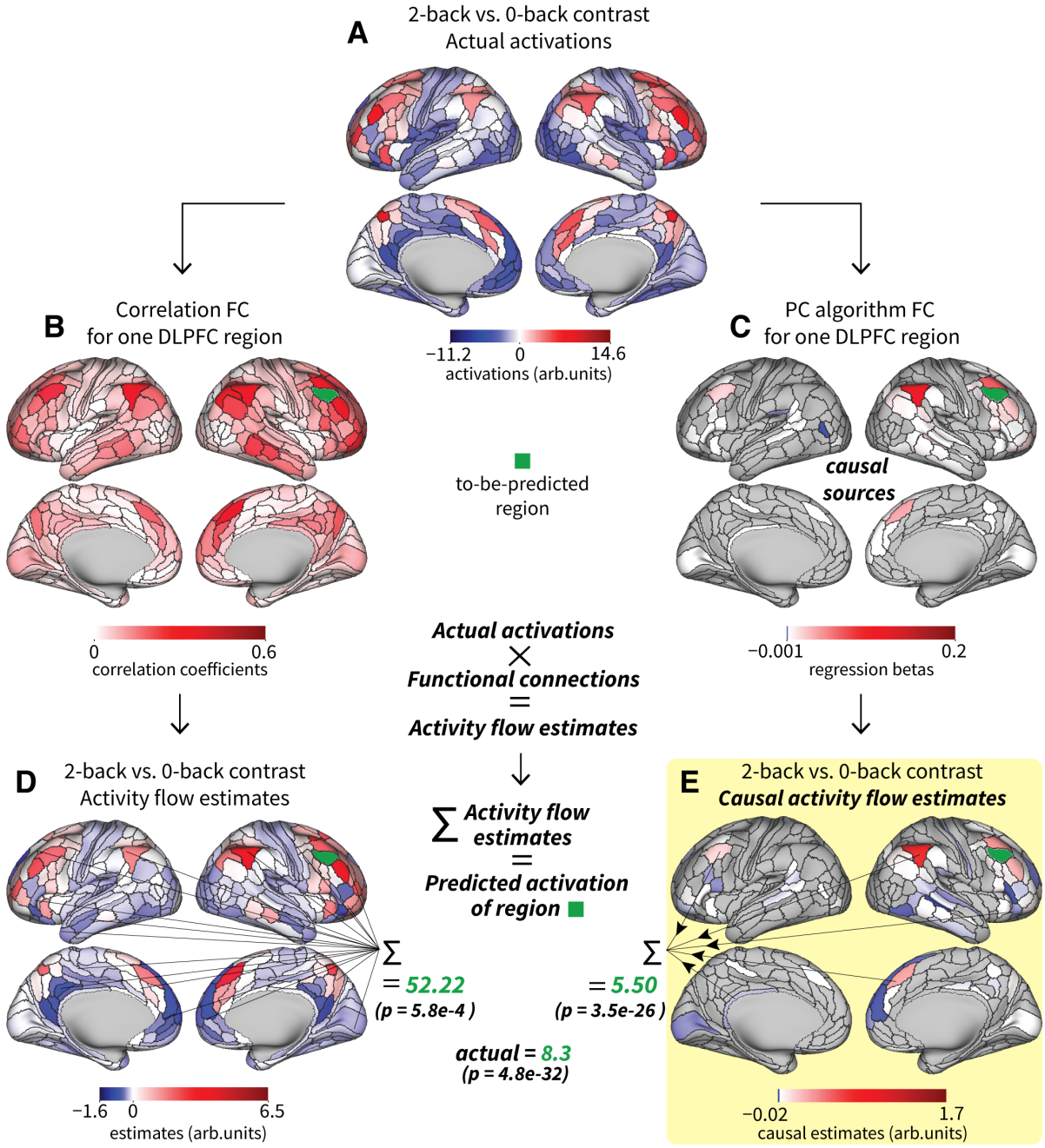

Figure 6 — Directed activity flow models provide more valid causal insight into the emergence of n-back cognitive responses in DLPFC. (A) 2-back vs. 0-back actual activation contrast, for each of the 360 Glasser cortical regions. (**B**) Correlation FC for one to-be-predicted region in the right hemisphere dorsolateral prefrontal cortex (green region, right DLPFC Area 8C), averaged across 176 participants. (**C**) PC algorithm directed FC from causal sources to the to-be-predicted right DLPFC Area 8C (green region). (**D**) Correlation-based activity flow estimates, resulting from multiplying each region's actual activation contrast (panel A) by its correlation functional connection (panel B), averaged across 176 participants. The region-wise sum of the activity flow estimates is the correlation-based actflow contrast prediction for the right DLPFC Area 8C; average predicted value across 176 participants, 52.22, *p*-value = 5.8e-4 for a two-sided *t*-test. (**E**) PC algorithm-based causal activity flow estimates, resulting from multiplying contrasts (panel A) by its PC causal functional connection (panel C). The region-wise sum of the causal activity flow estimates is the PC-based contrast prediction for the right DLPFC Area 8C;

average predicted value across 176 participants, 5.50, *p*-value = 3.5e-26 for a two-sided *t*-test. The actual activation contrast for the right DLPFC Area 8C is included for comparison, average across participants 8.3, *p*-value = 4.8e-32 for a two-sided *t*-test.

Table 1 — Contributing regions to the PC-causal DLPFC prediction. 21 regions (out of 95) with the strongest positive contributions (activity flow estimates) to predict right DLPFC Area 8C 2-back vs. 0-back working memory contrast. Values are averages across 176 participants. The collective removal of these regions from the actflow computation (simulated lesioning) makes the predicted contrast non-significant (average across participants 0.16, *p*-value>0.10 for a two-sided *t*-test). Region name and location from Glasser et al. (2016). See also Figure 6.

Region name	Location	Actual contrast activation	PC functional connection	Activity flow estimate
Area PFm Complex (right)	Inferior Parietal Cortex	8.8304	0.1962	1.6823
Area 8Av (r)	Dorsolateral Prefrontal Cortex	6.4347	0.1338	0.7911
Area 8BM (r)	Medial Prefrontal Cortex	9.9563	0.0758	0.7120
Area posterior 9-46v (r)	Dorsolateral Prefrontal Cortex	12.8559	0.0432	0.5672
Area 8C (left)	Dorsolateral Prefrontal Cortex	7.8699	0.0402	0.3433
Inferior 6-8 Transitional Area (r)	Dorsolateral Prefrontal Cortex	14.1699	0.0139	0.2203
Area anterior 47r (r)	Orbital and Polar Frontal Cortex	5.0893	0.0286	0.1349
Area IFJp (r)	Inferior Frontal Cortex	2.2545	0.0143	0.0873
Area posterior 10p (r)	Orbital and Polar Frontal Cortex	6.9349	0.0146	0.0864
Area PGs (r)	Inferior Parietal Cortex	5.3585	0.0171	0.0831
Area 44 (r)	Inferior Frontal Cortex	5.3350	0.0166	0.0827
Area 9 Posterior (r)	Dorsolateral Prefrontal Cortex	-1.5470	0.0154	0.0759
Area IntraParietal 1 (r)	Inferior Parietal Cortex	7.6205	0.0140	0.0754
Area IFSp (r)	Inferior Frontal Cortex	-0.0056	0.0240	0.0679
Anterior Ventral Insular Area (r)	Insular Cortex	10.8755	0.0042	0.0595
Superior 6-8 Transitional Area (r)	Dorsolateral Prefrontal Cortex	9.1950	0.0078	0.0582
Area anterior 9-46v (r)	Dorsolateral Prefrontal Cortex	10.3015	0.0077	0.0508
Area posterior 47r (l)	Inferior Frontal Cortex	4.4618	0.0031	0.0507
Area IFJa (r)	Inferior Frontal Cortex	0.5240	0.0132	0.0435
Area PGi (r)	Inferior Parietal Cortex	0.8612	0.0091	0.0404
Parieto-Occ. Sulcus Area 2 (r)	Posterior Cingulate Cortex	4.2712	0.0045	0.0348

Discussion

Using simulations and empirical fMRI we confirmed that FC methods grounded in causal principles can accurately recover properties of the underlying functional networks. These inferred networks can be used to construct actflow predictive models—empirically-constrained network simulations—that provide plausible causal explanations of the emergence of distributed cognitive brain activation effects.

Importantly, we illustrated the explanatory potential of activity flow models with a PC-based actflow model of right DLPFC (Area 8C), which accounts for the emergence of an established cognitive effect observed during increased working memory load (a statistically significant 2-back vs. 0-back contrast). We observed that the activation in right DLPFC Area 8C can be partially explained by the strong incoming flow of task-evoked activity from directly connected regions in inferior parietal, dorsolateral prefrontal and medial prefrontal cortices. In addition, we reported specific FC weights (as measures of communication between source regions and the to-be-predicted region) and activations (as measures of source regions sensitivity to the task conditions), to attain a fine-grained characterization of each source region's contribution to the predicted cognitive effect in right DLPFC Area 8C. Finally, we showed that simulated lesioning of a subset of the inferred source regions makes the predicted activation contrast non-significant, evidencing the functional relevance of these connections to explain the observed task-related effect. This approach goes beyond single reports of task-evoked activations of DLPFC during n-back conditions, or single reports of DLPFC connectivity patterns. Also, it is worth noting that our actflow-based mechanistic model of the DLPFC activation is derived from empirical brain data, which sets it apart from previous mechanistic models based on abstract computational neural networks (e.g., most artificial neural network models, since they are not directly constrained by empirical brain data).

Directed actflow models provide a basis for a new scientific paradigm for discovering network mechanisms underlying the emergence of neurocognitive phenomena. Most traditional explanations of cognitive phenomena are based on focal measures of neural responses to experimental interventions (Cabeza & Nyberg, 2000; Saxe et al., 2006). While this approach has been successful in establishing robust associations between brain regions and cognitive tasks (for example, DLPFC and working memory tasks or fusiform face area and face visual stimuli), these associations by themselves do not explain how task activity in a brain region emerges from underlying causal processes (e.g., brain network interactions). Patterns of activations have also been used in multivariate analyses to explain differences between task conditions (Kriegeskorte et al., 2008; Norman et al., 2006), but also fail to mechanistically explain how activations emerge from underlying causal processes. Another strategy to characterize associations between cognitive tasks and brain regions is to analyze changes in inter- or intra-region connectivity due to task manipulations (Gordon et al., 2014; Jolles et al., 2013; Vatansever et al., 2017). However, these studies do not typically assess the role of task-related neural activations, which are more clearly linked to cognition and behavior (e.g., activations in M1 are known to cause motor responses). Other explanations of cognitive effects are based on artificial neural network models

that try to reproduce empirically observed cognitive activity (Thomas & McClelland, 2008). However, most of these models are not constrained by empirical brain data (biological network architectures or signals), and thus provide only limited mechanistic insight. In contrast to these explanatory strategies, mechanistic actflow models successfully integrate empirically-derived connectivity models and empirical task activations to provide data-driven causal insight into network-supported cognitive processes in the human brain—as illustrated with our DLPFC example.

While the particular methods used here are illustrative of the proposed mechanistic actflow paradigm for explaining neurocognitive effects, there are many opportunities for further improvements. The limitations of the current study can perhaps be best illustrated by contrasting it with an idealized experiment not currently possible due to methodological limitations in neuroscience. Such an ideal study would observe all action potentials and local field potentials throughout the human brain in real time. This would contrast with the present study's use of fMRI, which has limited spatial resolution (2.0 mm voxels) and temporal resolution (720 ms time points) but whole-brain coverage that matches the ideal. FMRI detects blood-oxygen level dependent changes in MRI signals, which is an indirect measure of aggregate action potentials and local field potentials (Kahn et al., 2013; Logothetis et al., 2001; Shmuel & Leopold, 2008). Our data preprocessing procedures were designed to remove likely physiological artifacts (e.g., respiration, movement, and heart rate) present in fMRI data (see Methods).

The ideal study's perfect spatial coverage would be essential for reducing potential causal confounds. Given that fMRI has full spatial coverage, this benefit extended to the present study. However, the ideal study's perfect spatial resolution would allow the functional connectivity algorithms used in the present study to even better account for potential causal confounders, such as neural signals lost through averaging into observed voxel time series. The ideal study's perfect temporal resolution would present a challenge to the FC algorithms used here, since action potentials and downstream changes in local field potentials would occur with a temporal lag. Such lags would provide additional constraints on causal inferences not accounted for in the present study, but which are utilized in actflow studies utilizing higher temporal resolution methods (Mill et al., 2021). As mentioned above, another important component in the ideal study would be to follow the causal chain from stimulus to the cognitive phenomenon of interest (i.e., 2-back vs. 0-back activity increases in right DLPFC) to motor responses (behavior). This more comprehensive explanation of the DLPFC n-back effect is beyond the scope of the present study, but something close to this level of explanation—of a different set of neurocognitive phenomena—has been achieved in a recent actflow study (Ito et al., 2021). Finally, the ideal study would use stimulation and lesion approaches to fully verify the causal relevance of observed neural signals. While we utilized simulated lesioning to help verify the causal relevance of observed activations and connections, this illustrates an opportunity to use empirical stimulation and lesion approaches to verify the causal predictions made here.

We estimated causal networks using resting-state data, which has become a standard approach in neuroimaging and in actflow models. However, a recent study demonstrated improvements in actflow predictions when using task-state FC (relative to resting-state FC) (Cole et al., 2021). We

nonetheless chose to use resting-state FC here, given improvements in explanatory power with this approach. Specifically, successful actflow predictions using FC from a brain state (rest) other than the state of interest (e.g., 2-back task performance) provided evidence that our inferences were based on latent causal properties that generalize across brain states (McCormick et al., 2021). This is largely equivalent to estimating causal effects over structural connectivity networks, which (because structural connectivity generalizes over brain states) can be considered latent causal factors underlying dynamic neural processes. This approach contrasts with an account in which the identified causal interactions during the state of interest (e.g., 2-back task performance) would reveal no information regarding whether the identified mechanism generalizes beyond the observed state. Further, inferring causal networks using data from another state reduces the chance of overfitting to noise (Lever et al., 2016), which could potentially result in false causal inferences. An additional practical consideration is that FC estimation is substantially improved by including more time points, such that actflow predictions (and causal inferences) can be improved using the brain state(s) with the greatest amount of data (Cole et al., 2021; Sanchez-Romero & Cole, 2021). Using causal FC estimates from the state of interest nonetheless provides an opportunity for future research, given the possibility that some details of the actflow mechanisms generating cognitive effects were not observed with the present approach (e.g., task-specific DLPFC FC updates during the n-back task).

In our study, FC networks were inferred using the PC algorithm due to its strong causal principles, implementation simplicity, efficient running times, and adaptability (here we used linear conditional association tests, but nonlinear or nonparametric tests can be used if necessary). But other available directed FC methods could in principle be applied. Promising alternatives are an efficient implementation of the popular dynamical causal modelling (DCM) (Frässle et al., 2021), and artificial neural network modeling approaches such as the mesoscale individualized neurodynamic modeling (MINDy) (Singh et al., 2020) and current-based decomposition (CURBD) (Perich et al., 2020). These methods use different causal principles and estimation techniques other than the PC algorithm (and related Bayes networks methods (Ramsey et al., 2011, 2017)), and thus offer future opportunities to explore the robustness and diversity of mechanistic actflow explanations across diverse FC procedures.

The PC algorithm, as applied here, estimates contemporaneous functional associations from fMRI data, but for neural data acquired with higher temporal resolution, such as electroencephalography (EEG) or magnetoencephalography (MEG), it would be possible to apply temporal FC methods based on autoregressive processes (Amblard & Michel, 2013; Gilson et al., 2019; Malinsky & Spirtes, 2018; Moneta et al., 2011; Novelli et al., 2019; Runge, 2018; Shen et al., 2019) that identify temporally lagged functional interactions between brain regions. For example, Mill et al. (2021) have recently shown how temporally resolved FC networks can characterize with high precision the dynamics of the task-evoked causal activity flow that produce cognitive computations.

Tavor et al. (2016) presented an approach where task-evoked activations are predicted with resting-state connectivity maps and structural features as regressors, and recently Dohmatob et al. (2021) proposed a predictive model where resting-state spatial maps (e.g., default mode

network) are used as regressors to predict task activations—as opposed to linearly combining region-wise FC connections and activations as in actflow. However, while it is unclear that these predictive models offer tractable causal mechanistic accounts of cognitive task activations, their insight into the statistical associations between resting and task states may still provide complementary explanations relative to mechanistic actflow models.

Overall, the results presented here show that activity flow predictive models based on causal FC methods can accurately predict activations for a wide variety of brain regions and task conditions, offering the bases for a new explanatory paradigm with the flexibility and mechanistic properties to advance our understanding of network-based local and distributed cognitive computations in the human brain.

Acknowledgements

The authors acknowledge support by the US National Institutes of Health under awards R01 AG055556, R01 MH109520 and R01 EB022858-03. Data were provided, in part, by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: D. Van Essen and K. Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers The State University of New Jersey, for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies. The authors also thank Joseph D. Ramsey and Kevin Bui for support with the Tetrad software.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, *11*(1).

Amblard, P.-O., & Michel, O. J. (2013). The relation between Granger causality and directed information theory: A review. *Entropy*, *15*(1), 113–143. https://doi.org/10.3390/e15010113

Aquino, K. M., Fulcher, B. D., Parkes, L., Sabaroedin, K., & Fornito, A. (2020). Identifying and removing widespread signal deflections from fMRI data: Rethinking the global signal regression problem. *NeuroImage*, *212*, 116614. https://doi.org/10.1016/j.neuroimage.2020.116614

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., & Feldt, C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*(3), 47–53. https://www.jstor.org/stable/3002000

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Cabeza, R., & Nyberg, L. (2000). Imaging Cognition II: An Empirical Review of 275 PET and fMRI Studies. *Journal of Cognitive Neuroscience*, *12*(1), 1–47. https://doi.org/10.1162/08989290051137585

Carlson, S., Martinkauppi, S., Rämä, P., Salli, E., Korvenoja, A., & Aronen, H. J. (1998). Distribution of cortical activation during visuospatial n-back tasks as revealed by functional magnetic resonance imaging. *Cerebral Cortex*, *8*(8), 743–752. https://doi.org/10.1093/cercor/8.8.743

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Cole, M. W., Ito, T., Bassett, D. S., & Schultz, D. H. (2016). Activity flow over resting-state networks shapes cognitive task activations. *Nature Neuroscience*, *19*(12), 1718. https://doi.org/10.1038/nn.4406

Cole, M. W., Ito, T., Cocuzza, C., & Sanchez-Romero, R. (2021). The functional relevance of task-state functional connectivity. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1713-20.2021

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (2012). Global Connectivity of Prefrontal Cortex Predicts Cognitive Control and Intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999. https://doi.org/10.1523/JNEUROSCI.0536-12.2012

Colenbier, N., Van de Steen, F., Uddin, L. Q., Poldrack, R. A., Calhoun, V. D., & Marinazzo, D. (2020). Disambiguating the role of blood flow and global signal with partial information decomposition. *NeuroImage*, *213*, 116699. https://doi.org/10.1016/j.neuroimage.2020.116699

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, *15*(1), 3741–3782.

Constantinidis, C., Franowicz, M. N., & Goldman-Rakic, P. S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience*, *4*(3), 311–316. https://doi.org/10.1038/85179

Courtney, S. M., Petit, L., Maisog, J. M., Ungerleider, L. G., & Haxby, J. V. (1998). An Area Specialized for Spatial Working Memory in Human Frontal Cortex. *Science*, *279*(5355), 1347–1351. https://doi.org/10.1126/science.279.5355.1347

Dohmatob, E., Richard, H., Pinho, A. L., & Thirion, B. (2021). Brain topography beyond parcellations: Local gradients of functional maps. *NeuroImage*, *229*, 117706. https://doi.org/10.1016/j.neuroimage.2020.117706

Evangelista, N. D., O'Shea, A., Kraft, J. N., Hausman, H. K., Boutzoukas, E. M., Nissim, N. R., Albizu, A., Hardcastle, C., Van Etten, E. J., Bharadwaj, P. K., Smith, S. G., Song, H., Hishaw, G. A., DeKosky, S., Wu, S., Porges, E., Alexander, G. E., Marsiske, M., Cohen, R., & Woods, A. J. (2021). Independent Contributions of Dorsolateral Prefrontal Structure and Function to Working Memory in Healthy Older Adults. *Cerebral Cortex*, *31*(3), 1732–1743. https://doi.org/10.1093/cercor/bhaa322

Frässle, S., Harrison, S. J., Heinzle, J., Clementz, B. A., Tamminga, C. A., Sweeney, J. A., Gershon, E. S., Keshavan, M. S., Pearlson, G. D., Powers, A., & Stephan, K. E. (2021). Regression dynamic causal modeling for resting-state fMRI. *Human Brain Mapping*. https://doi.org/10.1002/hbm.25357

Fu, S., & Desmarais, M. C. (2010). Markov blanket based feature selection: A review of past decade. *Proceedings of the World Congress on Engineering*, *1*, 321–328.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, *36*, 193–202.

https://doi.org/10.1007/bf00344251

Gilson, M., Kouvaris, N. E., Deco, G., Mangin, J.-F., Poupon, C., Lefranc, S., Rivière, D., & Zamora-López, G. (2019). Network analysis of whole-brain fMRI dynamics: A new framework based on dynamic communicability. *NeuroImage*, *201*, 116007. https://doi.org/10.1016/j.neuroimage.2019.116007

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., & Jenkinson, M. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171. https://doi.org/10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gordon, E. M., Breeden, A. L., Bean, S. E., & Vaidya, C. J. (2014). Working memory-related changes in functional connectivity persist beyond task disengagement. *Human Brain Mapping*, *35*(3), 1004–1017. https://doi.org/10.1002/hbm.22230

Guyon, I., Aliferis, C., & Elisseeff, A. (2008). Causal feature selection. In H. Liu & H. Motoda (Eds.), *Computational methods of feature selection* (pp. 63–82). Chapman & Hall/CRC.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hanson, Stephen José, & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, *13*(3), 471–489. https://doi.org/10.1017/S0140525X00079760

Hearne, L., Mill, R., Keane, B., & Cole, M. (2020). Activity Flow Predictions Reveal the Role of Schizophrenia Network Abnormalities in Cognitive Activation and Behavioral Dysfunctions. *Biological Psychiatry*, *87*(9), S358. https://doi.org/10.1016/j.biopsych.2020.02.918

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Ito, T., Brincat, S. L., Siegel, M., Mill, R. D., He, B. J., Miller, E. K., Rotstein, H. G., & Cole, M. W. (2020). Task-evoked activity quenches neural correlations and variability across cortical areas. *PLOS Computational Biology*, *16*(8), e1007983. https://doi.org/10.1371/journal.pcbi.1007983

Ito, T., Hearne, L. J., & Cole, M. W. (2020). A cortical hierarchy of localized and distributed processes revealed via dissociation of task activations, connectivity changes, and intrinsic timescales. *NeuroImage*, *221*, 117141. https://doi.org/10.1016/j.neuroimage.2020.117141

Ito, T., Hearne, L., Mill, R., Cocuzza, C., & Cole, M. W. (2020). Discovering the computational relevance of brain network organization. *Trends in Cognitive Sciences*, *24*(1), 25–38.

Ito, T., Kulkarni, K. R., Schultz, D. H., Mill, R. D., Chen, R. H., Solomyak, L. I., & Cole, M. W. (2017). Cognitive task information is transferred between brain regions via resting-state network topology. *Nature Communications*, *8*(1), 1–14. https://doi.org/10.1038/s41467-017-01000-w

Ito, T., Yang, G. R., Laurent, P., Schultz, D. H., & Cole, M. W. (2021). Constructing neural network models from brain data reveals representational transformations underlying adaptive behavior. *BioRxiv*, 2020.12.24.424353. https://doi.org/10.1101/2020.12.24.424353

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006

Jolles, D. D., van Buchem, M. A., Crone, E. A., & Rombouts, S. A. (2013). Functional brain

connectivity at rest changes after working memory training. *Human Brain Mapping*, *34*(2), 396–406. https://doi.org/10.1002/hbm.21444

Kahn, I., Knoblich, U., Desai, M., Bernstein, J., Graybiel, A. M., Boyden, E. S., Buckner, R. L., & Moore, C. I. (2013). Optogenetic drive of neocortical pyramidal neurons generates fMRI signals that are correlated with spiking activity. *Brain Research*, *1511*, 33–45. https://doi.org/10.1016/j.brainres.2013.03.011

Keane, B. P., Barch, D. M., Mill, R., Silverstein, S. M., Krekelberg, B., & Cole, M. W. (2020). Brain network mechanisms of visual shape completion. *BioRxiv*. https://doi.org/10.1101/2020.08.03.233403

Kiiveri, H., Speed, T. P., & Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society*, *36*(1), 30–52. https://doi.org/10.1017/S1446788700027312

Kohonen, T. (1984). Classical Learning Systems. In *Self-Organization and Associative Memory* (Vol. 8). Springer.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kumar, S., Zomorrodi, R., Ghazala, Z., Goodman, M. S., Blumberger, D. M., Cheam, A., Fischer, C., Daskalakis, Z. J., Mulsant, B. H., & Pollock, B. G. (2017). Extent of dorsolateral prefrontal cortex plasticity and its association with working memory in patients with Alzheimer disease. *JAMA Psychiatry*, *74*(12), 1266–1274. https://doi.org/10.1001/jamapsychiatry.2017.3292

Li, J., Kong, R., Liegeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A. J., Sabuncu, M. R., Ge, T., & Yeo, B. T. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage*, *196*, 126–141. https://doi.org/10.1016/j.neuroimage.2019.04.016

Liu, T. T., Nalci, A., & Falahpour, M. (2017). The global signal in fMRI: Nuisance or Information? *NeuroImage*, *150*, 213–229. https://doi.org/10.1016/j.neuroimage.2017.02.036

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. https://doi.org/10.1038/35084005

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, *37*(6A), 3133–3164. https://doi.org/10.1214/09-AOS685

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, 23–47. http://proceedings.mlr.press/v92/malinsky18a

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McCormick, E. M., Arnemann, K. L., Ito, T., Hanson, S. J., & Cole, M. W. (2021). Latent functional connectivity underlying multiple brain states. *BioRxiv*, 2021.04.05.438534. https://doi.org/10.1101/2021.04.05.438534

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Meek, C. (1995). Causal inference and causal explanation with background knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 403–410.

Mill, R. D., Gordon, B. A., Balota, D. A., & Cole, M. W. (2020). Predicting dysfunctional age-related task activations from resting-state network alterations. *Neuroimage*, *221*, 117167. https://doi.org/10.1016/j.neuroimage.2020.117167

Mill, R. D., Hamilton, J. L., Winfield, E. C., Lalta, N., Chen, R. H., & Cole, M. W. (2021). Emergence of task information from dynamic network interactions in the human brain. *BioRxiv*, 2021.01.26.428276. https://doi.org/10.1101/2021.01.26.428276

Moneta, A., Chlaß, N., Entner, D., & Hoyer, P. (2011). Causal search in structural vector autoregressive models. *NIPS Mini-Symposium on Causality in Time Series*, 95–114. http://proceedings.mlr.press/v12/moneta11.html

Mumford, J. A., & Ramsey, J. D. (2014). Bayesian networks for fMRI: A primer. *Neuroimage*, *86*, 573–582. https://doi.org/10.1016/j.neuroimage.2013.10.020

Murphy, K., & Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage*, *154*, 169–173. https://doi.org/10.1016/j.neuroimage.2016.11.052

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Novelli, L., Wollstadt, P., Mediano, P., Wibral, M., & Lizier, J. T. (2019). Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, *3*(3), 827–847. https://doi.org/10.1162/netn_a_00092

Olson, C. R., & Hanson, C. R. (1990). Spatial Representation of the Body. In S.J. Hanson & C. R. Olson (Eds.), *Connectionist modeling and brain function: The developing interface* (pp. 193–254). MIT Press.

Panikratova, Y. R., Vlasova, R. M., Akhutina, T. V., Korneev, A. A., Sinitsyn, V. E., & Pechenkova, E. V. (2020). Functional connectivity of the dorsolateral prefrontal cortex contributes to different components of executive functions. *International Journal of Psychophysiology*, *151*, 70–79. https://doi.org/10.1016/j.ijpsycho.2020.02.013

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Perich, M. G., Arlt, C., Soares, S., Young, M. E., Mosher, C. P., Minxha, J., Carter, E., Rutishauser, U., Rudebeck, P. H., Harvey, C. D., & Rajan, K. (2020). Inferring brain-wide interactions using data-constrained recurrent neural network models. *BioRxiv*, 2020.12.18.423348. https://doi.org/10.1101/2020.12.18.423348

Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research*, *133*(1), 44–54. https://doi.org/10.1007/s002210000399

Ramsey, J. D. (2014). A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data. *ArXiv:1401.5031 [Cs, Stat]*. http://arxiv.org/abs/1401.5031

Ramsey, J. D. (2016). Improving accuracy and scalability of the pc algorithm by maximizing p-value. *ArXiv Preprint ArXiv:1610.00378*.

Ramsey, J. D., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). 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*(2), 121–129. https://doi.org/10.1007/s41060-016-0032-z

Ramsey, J. D., Hanson, S. J., & Glymour, C. (2011). Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. Simulation study. *NeuroImage*, *58*(3), 838–848. https://doi.org/10.1016/j.neuroimage.2011.06.068

Rashevsky, N. (1933). Outline of a physico-mathematical theory of excitation and inhibition. *Protoplasma*, *20*(1), 42–56. https://doi.org/10.1007/BF02674811

Reichenbach, H. (1956). *The direction of time* (Vol. 65). Univ of California Press.

Reid, A. T., Headley, D. B., Mill, R. D., Sanchez-Romero, R., Uddin, L. Q., Marinazzo, D., Lurie, D. J., Valdés-Sosa, P. A., Hanson, S. J., Biswal, B. B., Calhoun, V., Poldrack, R. A., & Cole, M. W. (2019). Advancing functional connectivity research from association to causation. *Nature Neuroscience*, 1–10. https://doi.org/10.1038/s41593-019-0510-4

Reineberg, A. E., & Banich, M. T. (2016). Functional connectivity at rest is sensitive to individual differences in executive function: A network analysis. *Human Brain Mapping*, *37*(8), 2959–2975. https://doi.org/10.1002/hbm.23219

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386. https://psycnet.apa.org/doi/10.1037/h0042519

Rowe, J. B., & Passingham, R. E. (2001). Working Memory for Location and Time: Activity in Prefrontal Area 46 Relates to Selection Rather than Maintenance in Memory. *NeuroImage*, *14*(1), 77–86. https://doi.org/10.1006/nimg.2001.0784

Rowe, J. B., Toni, I., Josephs, O., Frackowiak, R. S. J., & Passingham, R. E. (2000). The Prefrontal Cortex: Response Selection or Maintenance Within Working Memory? *Science*, *288*(5471), 1656–1660. https://doi.org/10.1126/science.288.5471.1656

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *1*(45–76), 26.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

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

Sanchez-Romero, R., & Cole, M. W. (2021). Combining multiple functional connectivity methods to improve causal inferences. *Journal of Cognitive Neuroscience*, *33*(2), 180–194. https://doi.org/10.1162/jocn_a_01580

Sanchez-Romero, R., Ramsey, J. D., Zhang, K., Glymour, M. R., Huang, B., & Glymour, C. (2019). Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, *3*(2), 274–306. https://doi.org/10.1162/netn_a_00061

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062

Shen, Y., Giannakis, G. B., & Baingana, B. (2019). Nonlinear structural vector autoregressive models with application to directed brain networks. *IEEE Transactions on Signal Processing*, *67*(20), 5325–5339. https://doi.org/10.1109/TSP.2019.2940122

Sherwood, M. S., Weisend, M. P., Kane, J. H., & Parker, J. G. (2016). Combining Real-Time fMRI Neurofeedback Training of the DLPFC with N-Back Practice Results in Neuroplastic Effects Confined to the Neurofeedback Target Region. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00138

Shmuel, A., & Leopold, D. A. (2008). Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: Implications for functional connectivity at rest. *Human Brain Mapping*, *29*(7), 751–761. https://doi.org/10.1002/hbm.20580

Singh, M. F., Braver, T. S., Cole, M. W., & Ching, S. (2020). Estimation and validation of individualized dynamic brain models with resting state fMRI. *NeuroImage*, *221*, 117046.

https://doi.org/10.1016/j.neuroimage.2020.117046

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., & Harms, M. P. (2013). Resting-state fMRI in the human connectome project. *Neuroimage*, *80*, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, *9*(1), 62–72. https://doi.org/10.1177/089443939100900106

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (Second edition). The MIT Press.

Tavor, I., Jones, O. P., Mars, R. B., Smith, S. M., Behrens, T. E., & Jbabdi, S. (2016). Task-free MRI predicts individual differences in brain activity during task performance. *Science*, *352*(6282), 216–220. https://doi.org/10.1126/science.aad8127

Thomas, M. S. C., & McClelland, J. L. (2008). Connectionist models of cognition. In *The Cambridge handbook of computational psychology* (pp. 23–58). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.005

Ugurbil, K., Xu, J., Auerbach, E. J., Moeller, S., Vu, A. T., Duarte-Carvajalino, J. M., Lenglet, C., Wu, X., Schmitter, S., & Van de Moortele, P. F. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage*, *80*, 80–104. https://doi.org/10.1016/j.neuroimage.2013.05.012

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Consortium, W.-M. H. (2013). The WU-Minn human connectome project: An overview. *Neuroimage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Vatansever, D., Manktelow, A. E., Sahakian, B. J., Menon, D. K., & Stamatakis, E. A. (2017). Angular default mode network connectivity across working memory load. *Human Brain Mapping*, *38*(1), 41–52. https://doi.org/10.1002/hbm.23341

Von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, *14*(2), 85–100. https://doi.org/10.1007/BF00288907

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: *Cognitive, Affective, & Behavioral Neuroscience*, *3*(4), 255–274. https://doi.org/10.3758/CABN.3.4.255

Weichwald, S., & Peters, J. (2021). Causality in Cognitive Neuroscience: Concepts, Challenges, and Distributional Robustness. *Journal of Cognitive Neuroscience*, *33*(2), 226–247. https://doi.org/10.1162/jocn_a_01623

Widrow, B. (1962). Generalization and Information Storage in Networks of ADALINE Neurons. Self Organizing Systems. In M. C. Yovitz, G. T. Jacobi, & G. Goldstein (Eds.), *Self-Organizing Systems, symposium proceedings* (pp. 435–461). Spartan Books.

Woodcock, E. A., Anand, C., Khatib, D., Diwadkar, V. A., & Stanley, J. A. (2018). Working Memory Modulates Glutamate Levels in the Dorsolateral Prefrontal Cortex during 1H fMRS. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/fpsyt.2018.00066

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based Conditional Independence Test and Application in Causal Discovery. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 804–813. https://dl.acm.org/doi/10.5555/3020548.3020641