Whole-Brain Causal Discovery Using fMRI

Fahimeh Arab[1], AmirEmad Ghassami[2], Hamidreza Jamalabadi[3], Megan A. K. Peters[4,5,6], and Erfan Nozari[1,7,8,*]

[1]Department of Electrical and Computer Engineering, University of California, Riverside, USA
[2]Department of Mathematics and Statistics, Boston University, USA
[3]Department of Psychiatry and Psychotherapy, Phillips University of Marburg, Germany
[4]Department of Cognitive Sciences, University of California, Irvine, USA
[5]Center for the Neurobiology of Learning & Memory, University of California, Irvine, USA
[6]Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research, Canada
[7]Department of Mechanical Engineering, University of California, Riverside, USA
[8]Department of Bioengineering, University of California, Riverside, USA
[*]Corresponding author (email: erfan.nozari@ucr.edu)

Abstract

Despite significant research, discovering causal relationships from fMRI remains a challenge. Popular methods such as Granger Causality and Dynamic Causal Modeling fall short in handling contemporaneous effects and latent common causes. Methods from causal structure learning literature can address these limitations but often scale poorly with network size and need acyclicity. In this study, we first provide a taxonomy of existing methods and compare their accuracy and efficiency on simulated fMRI from simple topologies. This analysis demonstrates a pressing need for more accurate and scalable methods, motivating the design of Causal discovery for Large-scale Low-resolution Time-series with Feedback (CaLLTiF). CaLLTiF is a constraint-based method that uses conditional independence between contemporaneous and lagged variables to extract causal relationships. On simulated fMRI from the macaque connectome, CaLLTiF achieves significantly higher accuracy and scalability than all tested alternatives. From resting-state human fMRI, CaLLTiF learns causal connectomes that are highly consistent across individuals, show clear top-down flow of causal effect from attention and default mode to sensorimotor networks, exhibit Euclidean distance-dependence in causal interactions, and are highly dominated by contemporaneous effects. Overall, this work takes a major step in enhancing causal discovery from whole-brain fMRI and defines a new standard for future investigations.

Keywords: fMRI, causal discovery, brain networks, statistical algorithms, cognitive neuroscience

Introduction

A major step in the global drive for understanding the brain (Adams et al., 2020; Amunts et al., 2016; Jorgenson et al., 2015; Okano et al., 2015; Poo et al., 2016) is to move beyond correlations and understand the causal relationships among internal and external factors – a process often referred to as *causal discovery* (Assaad et al., 2022; Glymour et al., 2019). When possible, causal discovery can be greatly simplified by intervening in one variable and observing the effect in others. However, such interventions are often costly and/or infeasible, necessitating the significantly more challenging task of causal discovery from purely observational data.

A particularly rich set of observational data for the brain comes from functional MRI (fMRI) (Goense and Logothetis, 2008; Winder AT and PJ, 2017). The whole-brain coverage allowed by fMRI is valuable for causal discovery not only because it allows for purely data-driven and unbiased discovery of potentially unexpected causal relationships (Bressler and Menon, 2010; Bullmore and Sporns, 2012; Fornito and Bullmore, 2015), but also because of the great extent to which the presence of unobserved variables can complicate delineating causal adjacencies and orientations (Entner and Hoyer, 2010; Gerhardus and Runge, 2020; Hasan et al., 2023; Malinsky and Spirtes, 2018). Nevertheless, many characteristics of fMRI also make causal discovery

challenging, including its large dimensionality, low temporal resolution, and indirect reflection of underlying neural processes (Ramsey et al., 2010).

This has motivated a large and growing body of literature on causal discovery from fMRI. A common approach for causal discovery using neuroimaging and neurophysiology data is Granger Causality (GC) (Seth et al., 2015). GC has a long history in neuroscience (Barnett and Seth, 2014; Seth et al., 2015), but also has well-known limitations, including its lack of ability to account for contemporaneous causal relationships and the presence of latent nodes. The former is particularly important for fMRI. The temporal resolution in fMRI is typically within a few hundred milliseconds to several seconds (Huettel et al., 2009), which is about one order of magnitude slower than the time that it takes for neural signals to travel across the brain (Nunez and Srinivasan, 2006; Sutton and Begleiter, 1979; Sutton et al., 1965). Therefore, from one fMRI sample to the next, there is enough time for causal effects to flow between almost all pairs of nodes in the network (cf. a related in-depth discussion in (Nozari et al., 2019, Appendix A)). Such fast sub-TR interactions demonstrate themselves as causal effects that appear to be "contemporaneous" and can even be cyclic, making causal discovery significantly more challenging (Supplementary Note 1). Similar to GC, Dynamic Causal Modeling (DCM) has also been widely used with fMRI data (Friston et al., 2003, 2019, 2014; Stephan and Roebroeck, 2012) and fundamentally relies on the temporal order of a generative dynamical model to infer causation from correlations, making it similarly unable to account for contemporaneous causal relationships (Friston et al., 2013; Friston, 2011; Logothetis, 2008).

Discovering causal relationships without reliance on time has been the subject of extensive research in the causal inference literature (Glymour et al., 2019; Pearl, 1988, 2009a,b; Spirtes et al., 2000; Spirtes and Zhang, 2016). A wide range of algorithmic solutions have been proposed (Chickering, 2002a; Glymour et al., 2019; Henry and Gates, 2017; Meek, 1995, 1997; Pearl, 2009a; Ramsey et al., 2010; Shimizu et al., 2006; Smith et al., 2011; Spirtes and Glymour, 1991; Spirtes et al., 1995), which are often classified based on their methodology into constraint-based (Dawid, 1979; Pearl, 1988, 2009b), noise-based (Shimizu, 2014; Shimizu et al., 2006), and score-based (Chickering, 2002b; Heckerman et al., 1995). Nevertheless, which of these algorithms are suitable for whole-brain fMRI causal discovery and how they compare against each other in terms of accuracy and scalability have remained largely unknown.

In this study, we first discuss and compare existing causal discovery algorithms for their suitability for whole-brain fMRI, demonstrate a large gap between what causal discovery for fMRI needs and what existing algorithms can achieve, propose CaLLTiF to address this gap, and demonstrate its higher accuracy and scalability on synthetic and real fMRI.

Results

A Taxonomy of Causal Discovery for Whole-Brain fMRI

A vast array of algorithmic solutions exist for learning causal graphs from observational data, but not all are suitable for fMRI data. We selected a subset of state-of-the-art algorithms suitable for whole-brain fMRI data based on four criteria: (1) ability to learn cycles, (2) ability to learn contemporaneous effects, (3) assuming complete coverage of relevant variables in observed data, and (4) linearity (see Discussions). Table 1 shows several state-of-the-art methods that satisfy criteria (1)-(4). Multivariate Granger Causality (MVGC) (Barnett and Seth, 2014; Granger, 1969) does not satisfy criteria (2), but we still included it in our subsequent analyses due to its popularity in neuroscience (Ding et al., 2006; Goebel et al., 2003; Liao et al., 2010; Roebroeck et al., 2005). On the other hand, we excluded LiNG (Ramsey et al., 2018) from further analysis since it is considered by its proposers as generally inferior to the hybrid FASK algorithm (Sanchez-Romero et al., 2019). We also chose FASK for implementation over GANGO (Rawls et al., 2022), a similarly hybrid method with the additional caveat of not having a unified publicly available code distribution.

We compared the accuracy of the resulting list of algorithms (MVGC, PCMCI, PCMCI$^+$, VARLiNGAM, DYNOTEARS, FASK, and DGlearn) using simulated fMRI data from a benchmark of simple (5-10 nodes) networks introduced in (Sanchez-Romero et al., 2019). The ground truth graphs are shown in Figure 1a, and details on the fMRI time series generation for each node in these graphs are provided under Methods. To evaluate the success of each algorithm, we treated the causal discovery problem as a binary classification problem for each directed edge and calculated the resulting F1 score, both for the directed graphs as well as their undirected skeletons (see Methods for details). Figure 1b illustrates the distribution of F1 scores

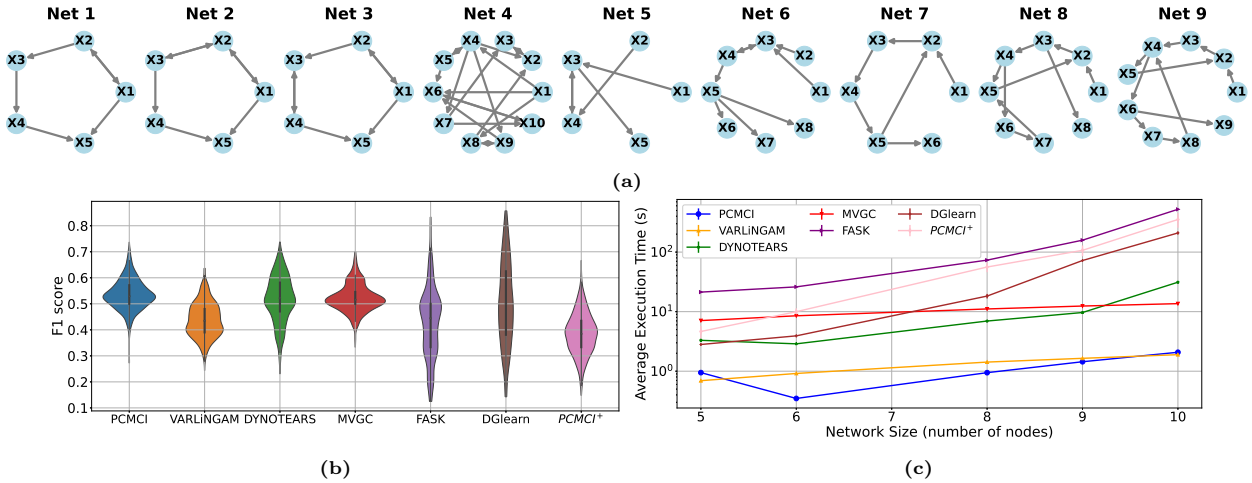


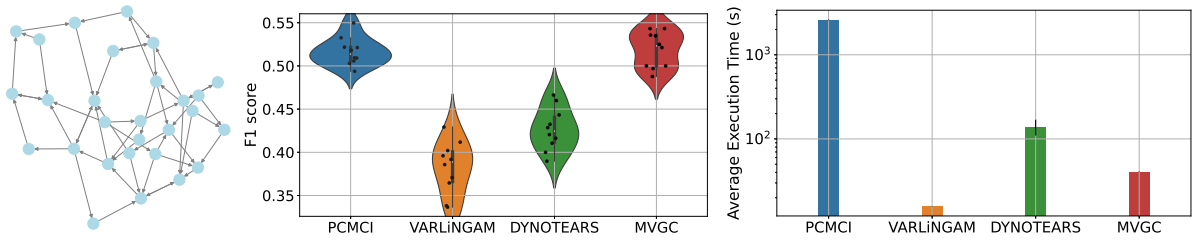
Figure 1: Results of comparing several state-of-the-art causal discovery algorithms over a benchmark of simulated fMRI (Sanchez-Romero et al., 2019) generated from simple networks with 5-10 nodes each. (a) Ground truth graphs of the simple networks in the benchmark. Despite all being small-scale, the graphs vary widely in their density, number of cycles, etc. **(b)** F1 scores of seven state-of-the-art algorithms (six from Table 1 and MVGC) for correctly identifying the full (directed) graphs. All methods are evaluated using optimized values of their respective hyperparameters (see Methods). The benchmark data includes 60 repetitions of fMRI data from each of the 9 graphs, so each violin plot is based on 540 F1 score samples. The PCMCI method achieves the highest median F1 score, both directed and undirected (see Supplementary Figure 3). **(c)** The mean execution time (averaged over all 60 repetitions) of different algorithms as a function of network size. Note the logarithmic scaling of the vertical axis. Error bars, though hardly visible, show 1 s.e.m.

for all algorithms, combined across nine simple networks. The results show that the PCMCI algorithm achieved significantly higher median F1 score compared to all other algorithms over the directed graphs (all Cohen's d > 0.23 and $p < 10^{-4}$, pairwise one-sided Wilcoxon signed-rank test) and compared to all but DYNOTEARS over the undirected skeletons (Supplementary Figure 3, all Cohen's d > 0.44 and $p < 10^{-29}$, pairwise one-sided Wilcoxon signed-rank test. Also see Supplementary Figures 4 and 5 for precision and recall). The PCMCI algorithm also has the smallest computational complexity on simple networks, as seen from Figure 1c. Furthermore, our results indicate that FASK, DGlearn, and PCMCI$^+$ (at their best values of hyperparameters) do not scale well with network size, forcing us to exclude them from further analysis as we move on to larger networks (see Supplementary Figures 13, 14, and 15).

Next, we compared the remaining four algorithms (PCMCI, MVGC, DYNOTEARS, and VARLiNGAM) on a larger-scale, more realistic simulated benchmark. The graph shown in Figure 2a, called the 'Small-Degree Macaque' network, consists of a complete macaque anatomical connectome with 28 nodes and 52 directed edges (Sanchez-Romero et al., 2019) but the generative model used to simulate fMRI data from this graph remains the same (see Methods for details). The distributions of F1 scores are shown in Figure 2b. PCMCI and MVGC achieved very similar success in learning both the full graph and its undirected skeleton, while significantly outperforming DYNOTEARS and VARLiNGAM. A similar result is obtained when comparing adjacency F1 scores for detecting the network's undirected skeleton (Supplementary Figure 17, also see Supplementary Figures 18 and 19 for precision and recall). As far as execution time is concerned, however, MVGC showed a significant advantage over PCMCI (Figure 2c). Therefore, despite its simplistic nature, MVGC was found most successful in causal discovery from *medium-sized* simulated macaque fMRI data (but

Table 1: List of causal discovery methods suitable for use with whole-brain fMRI, divided by methodological category (constraint-, noise-, and score-based). All these methods (1) allow for cycles, (2) allow for contemporaneous effects, (3) assume complete coverage of relevant variables in observed data, and (4) learn linear relationships. The FASK algorithm is fundamentally hybrid and therefore listed as both constraint-based and noise-based.

Type \ Category	Constraint-based	Noise-based	Score-based
Time-series	PCMCI (Runge et al., 2019), PCMCI$^+$ (Runge, 2020)	VARLiNGAM (Hyvärinen et al., 2010)	DYNOTEARS (Pamfil et al., 2020)
Cross-sectional with cycles	FASK (Sanchez-Romero et al., 2019)	FASK (Sanchez-Romero et al., 2019), LiNG (Lacerda et al., 2008)	DGlearn (Ghassami et al., 2020)



(a) **(b)** **(c)**

Figure 2: Comparing the scalable subset of algorithms from Figure 1 over simulated fMRI data from the Small Degree Macaque benchmark (Sanchez-Romero et al., 2019). (a) ground-truth Small Degree Macaque network. **(b)** F1 scores of identifying the full (directed) graph. Each distribution consists of 10 F1 scores calculated based on 10 repetitions of simulated data from the same underlying graph. **(c)** Mean execution times for each method (error bars show one standard deviation).

also see Figure 3).

CaLLTiF: A New Causal Discovery Algorithm for Whole-Brain fMRI

The best-performing algorithms on Small-Degree Macaque, i.e., PCMCI and MVGC, suffer from three main drawbacks: (1) poor scalability (only for PCMCI), (2) inability to learn directed contemporaneous effects (PCMCI only learns undirected contemporaneous effects while MVGC learns none), and (3) having sparsity-controlling hyperparameters that are subjectively selected in the absence of ground-truth graphs. In this section, we describe the design of a new algorithm based on PCMCI that mitigates these drawbacks and demonstrate its superior performance over existing methods.

Our first modification to PCMCI is with regard to scalability and computational complexity. The computational complexity of PCMCI depends heavily on the value of its 'PC Alpha' hyperparameter, which controls the sparsity of the set of potential common causes on which the algorithm conditions when checking the conditional independence of each pair of nodes (Supplementary Figure 16a). Higher values of PC Alpha make these sets denser and accordingly decrease statistical power in the subsequent conditional independence tests, *ultimately conditioning on all other nodes (and all of their lags) when PC Alpha = 1.* Nevertheless, interestingly, our experiments on the Small-Degree Macaque data show that the maximum achievable accuracy of PCMCI (i.e., F1 score maximized over Alpha Level for each fixed value of PC Alpha) monotonically increases with PC Alpha, reaching its maximum at PC Alpha = 1 (Supplementary Figures 16b and 16c). Therefore, while this may seem to cause a trade-off between accuracy and scalability, it is in fact an opportunity for maximizing both. At PC Alpha = 1, the PC part of PCMCI (a.k.a. the S1 algorithm in (Runge et al., 2019)) is theoretically guaranteed to return a complete conditioning set for all pairs of nodes, and can thus be skipped entirely. The PC part is further responsible for the poor scalability of PCMCI. Thus its removal significantly improves the computational efficiency of the resulting algorithm *without compromising accuracy* (cf. Discussions for a potential explanation of why conditioning on all other nodes may improve accuracy despite lowering statistical power).

Our second modification addresses the lack of directed contemporaneous causal effects (see Introduction for why these effects are particularly important in fMRI). By default, MVGC returns no contemporaneous edges and PCMCI returns $\circ\!\!-\!\!\circ$ ones which only indicate the presence of significant partial correlations but does not resolve between $\rightarrow$, $\leftarrow$, or $\leftrightarrows$. However, we know from decades of tract tracing studies that reciprocal connections are significantly more common than unidirectional connections in the primate brain (Felleman and Van Essen, 1991; Markov et al., 2014; Tigges et al., 1973). Therefore, we replace all $\circ\!\!-\!\!\circ$ edges returned by PCMCI by the more likely choice of $\leftrightarrows$. The only exception comes from (the often minority of) pairs of nodes that have a *lagged* directed edge between them, in which case we leave the direction of the contemporaneous effect between them the same as their lagged effect.

Figure 3 shows how the resulting CaLLTiF algorithm performs on a synthetic fMRI dataset generated from the significantly larger full macaque structural connectome with 91 nodes and 1,615 ground-truth edges (Figure 3a, see also Methods). CaLLTiF has a significantly higher F1 score compared to PCMCI, VARLiNGAM, DYNOTEARS, and MVGC in its discovery of the complete directed graph (Figure 3b, all

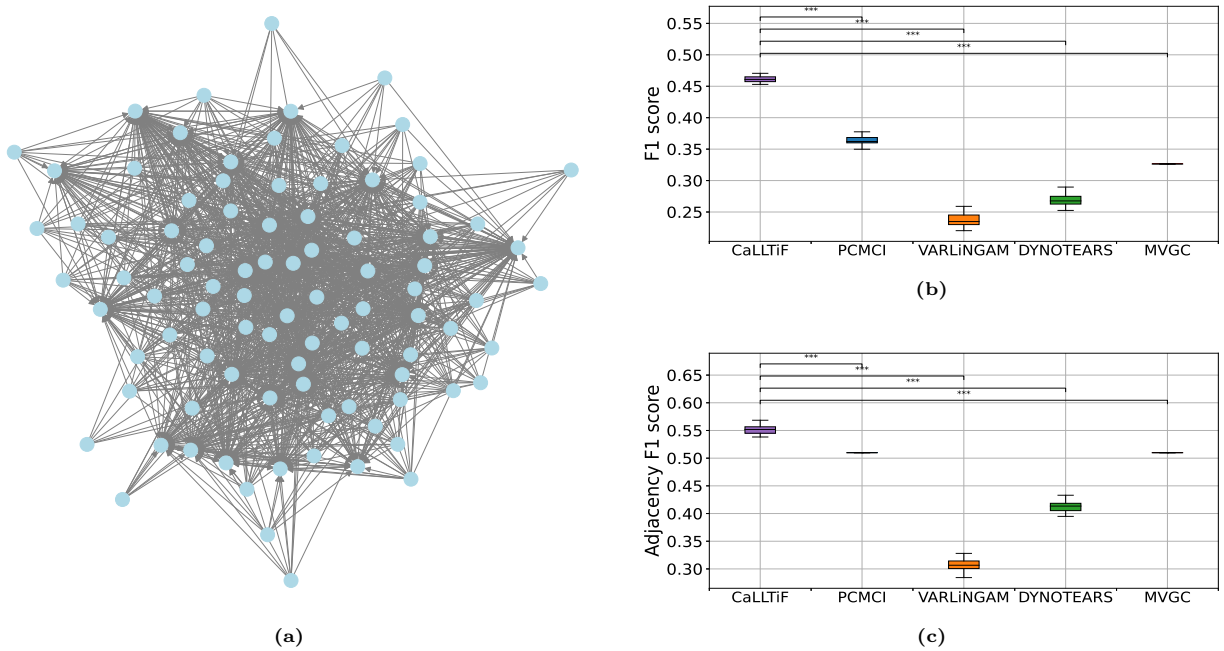


Figure 3: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. (a) ground-truth Full Macaque network. **(b,c)** Distributions of F1 scores for CaLLTiF and state-of-the-art alternatives in discovering the full (directed) graph (b) and its undirected adjacency skeleton (c). For all repetitions, the best performance of MVGC occurs at $\alpha = 0.5$ which returns a complete graph, hence the point distributions for MVGC. *** denotes $p < 0.001$. All statistical comparisons are performed using a one-sided Wilcoxon signed-rank test. In all boxplots, the center line represents the median, the box spans the interquartile range (IQR), and the whiskers extend up to 1.5 times the IQR from the box limits.

Cohen's $d > 15$, $p < 10^{-6}$, one-sided Wilcoxon signed-rank test) as well as its undirected adjacency skeleton (Figure 3c, all Cohen's $d > 7$, $p < 10^{-6}$, one-sided Wilcoxon signed-rank test). Precisions and recalls are also shown in Supplementary Figures 25 and 26. We also compared CaLLTiF (and PCMCI) against a middle-ground 'Mixed-PCMCI' variant where the $\circ\!-\!\circ$ edges returned by PCMCI are used only in the computation of adjacency F1 score (Supplementary Figures 27-29, see also Methods). Mixed-PCMCI benefits from contemporaneous effects as much as CaLLTiF in terms of adjacency F1 score, but not so in terms of full F1 score, further motivating the inclusion of directed contemporaneous connections as done in CaLLTiF. Detailed performances of all compared algorithms are provided in Supplementary Figures 28-35.

Finally, the third aspect in which CaLLTiF departs from PCMCI is the choice of sparsity-controlling hyperparameter 'Alpha Level'. Most, if not all, algorithms for causal discovery have at least one hyperparameter (often a threshold) that controls the sparsity of the resulting graphs. Different from PC Alpha described earlier, Alpha Level in PCMCI is the standard type-I error bound in determining statistical significance in *each* partial correlation test (cf. Supplementary Figure 16). By default, Alpha Level is selected subjectively, based on domain knowledge and expected level of sparsity. However, in CaLLTiF, we select Alpha Level objectively based on a novel method for correction for multiple comparisons (see Methods) that occur when collapsing a time-series graph over lagged variables into a final summary graph. This step is critical, particularly in the absence of ground-truth connectivity, to ensure that we have statistical confidence in every edge of the final summary graph returned by CaLLTiF.

In summary, CaLLTiF starts by constructing an extended time-lagged graph among all the variables $X_i(t-\tau), i = 1, \ldots, n$ and all lags $\tau = 0, 1, \ldots, \tau_{\max}$. To establish a causal link between any pair of variables $X_i(t-\tau)$ and $X_j(t)$, CaLLTiF performs a conditional independence test (using linear partial correlation) between $X_i(t-\tau)$ and $X_j(t)$, conditioned on all other *lagged* variables $(X_k(t-s), s = 1, \ldots, \tau_{\max})$. A causal link is established if the null hypothesis of conditional independence is rejected at a significance threshold 'Alpha Level'. By default, 'Alpha Level' is selected based on CaLLTiF's type I error control, but it can also be optimized in simulated data using ground-truth knowledge. If $\tau > 0$, the direction of the edge is clearly

$X_i(t-\tau) \to X_j(t)$. When $\tau = 0$, CaLLTiF returns $X_i(t) \leftrightarrows X_j(t)$ if no other edges exist between X_i and X_j at higher lags, and places the edge(s) consistent with the corresponding lagged direction(s) otherwise. Finally, the extended time-lagged graph is collapsed into a summary graph by taking an OR operation for each edge across all lags (cf. Methods for details).

Causal Discovery from Resting-State Human fMRI

We next applied CaLLTiF on resting-state fMRI from 200 subjects from the Human Connectome Project (HCP) (See Methods). Each scan from each subject was parcellated into 100 cortical and 16 subcortical regions. CaLLTiF was then performed on all four resting-state scans for each subject, resulting in one causal graph per individual.

Learned causal graphs are highly consistent across subjects. Despite individual differences, a remarkably common causal connectome emerged across subjects. Figure 4a shows the average causal graph among the subjects and Figure 4b shows the intersection graph that contained the edges *common across all subjects*. Due to the binary nature of individual graphs, the former can also be viewed as a matrix of probabilities, where entry (i, j) shows the probability of region i causing region j across all subjects. As a result of the significant commonalities that exist in the causal graphs among subjects, the average graph has a bimodal distribution, with the vast majority of average weights being close to either 0 or 1. These extreme values of average weights can also be seen as a measure of the confidence of the algorithm in the presence or lack of most edges, and have a clear contrast with the weights of the average of randomized surrogate graphs generated *independently* across subjects (Supplementary Figure 42). In the absence of a ground truth causal connectivity for direct comparison, such strong commonalities among subjects serve as an alternative measure of validation and provide insights into the general patterns and characteristics of the causal relationships in a resting brain.

Nodal centralities also show strong consistency among the subjects. Figure 4c shows the nodal degrees for all subjects (gray lines) as well as its mean across subjects (black line, also depicted in Figure 4d). Statistically significant differences exist between the degree distributions of many pairs of nodes (about 90% of the pairs have $p < 0.001$, two-sided Wilcoxon signed-rank test, computed between nodal degrees of each pair of parcels), while significant correlations exist between nodal degrees for all pairs of subjects (all pairs have $0.56 \leq r \leq 0.96$, $p < 10^{-10}$, Pearson Correlation Coefficient, computed based on the nodal degrees of each pair of subjects separately). Similar consistency for in-degree, out-degree, betweenness, and eigenvector centralities can be observed among subjects (Supplementary Figures 43-46). Consistently across subjects, medial ventral attention regions, cingulate cortices, and lateral primary sensorimotor areas show particularly low nodal degrees across both hemispheres, whereas bilateral default mode areas, particularly the left ventromedial prefrontal cortex, show notably strong nodal degrees. Bilateral anterior thalami are particularly causally connected compared to other subcortical regions, even though subcortical areas have significantly lower degrees than cortical areas in general, with bilateral posterior thalami, nuclei accumbens, and globus pallidi showing the least causal connections across the whole brain at rest.

Causal graphs are also sparser and more consistent across subjects compared to functional connectivity. A major motivation for building causal connectomes is the removal of spurious connections in functional connectivity (FC) profiles that reflect mere correlation but no causation. For causal graphs learned by CaLLTiF, we indeed observed significantly lower edge density compared to FC graphs (see Methods for details on the computation of FC graphs) (Supplementary Figures 41a,41b, no overlap existing between the support of the two distributions). In fact, FC graphs included approximately 95% of CaLLTiF's discovered causal edges (Supplementary Figure 41c), while only about half of all functional connectivity edges are also causal (Supplementary Figure 41d). Interestingly, among the approximately 5% of causal edges that were not in the FC graphs, the majority came from non-zero lags. This is remarkable, given that causal edges from non-zero lags are significantly fewer in general (cf. Figure 6a), but are fundamentally not discoverable by FC which only measures contemporaneous co-fluctuations. Moreover, causal connectomes are significantly more consistent across subjects compared to FC connectomes (Supplementary Figure 41e, Cohen'd > 2, $p < 0.001$, one-sided Wilcoxon signed-rank test), further reinforcing the expectation that causal edges are "pruned" and more reliable compared to functional edges.

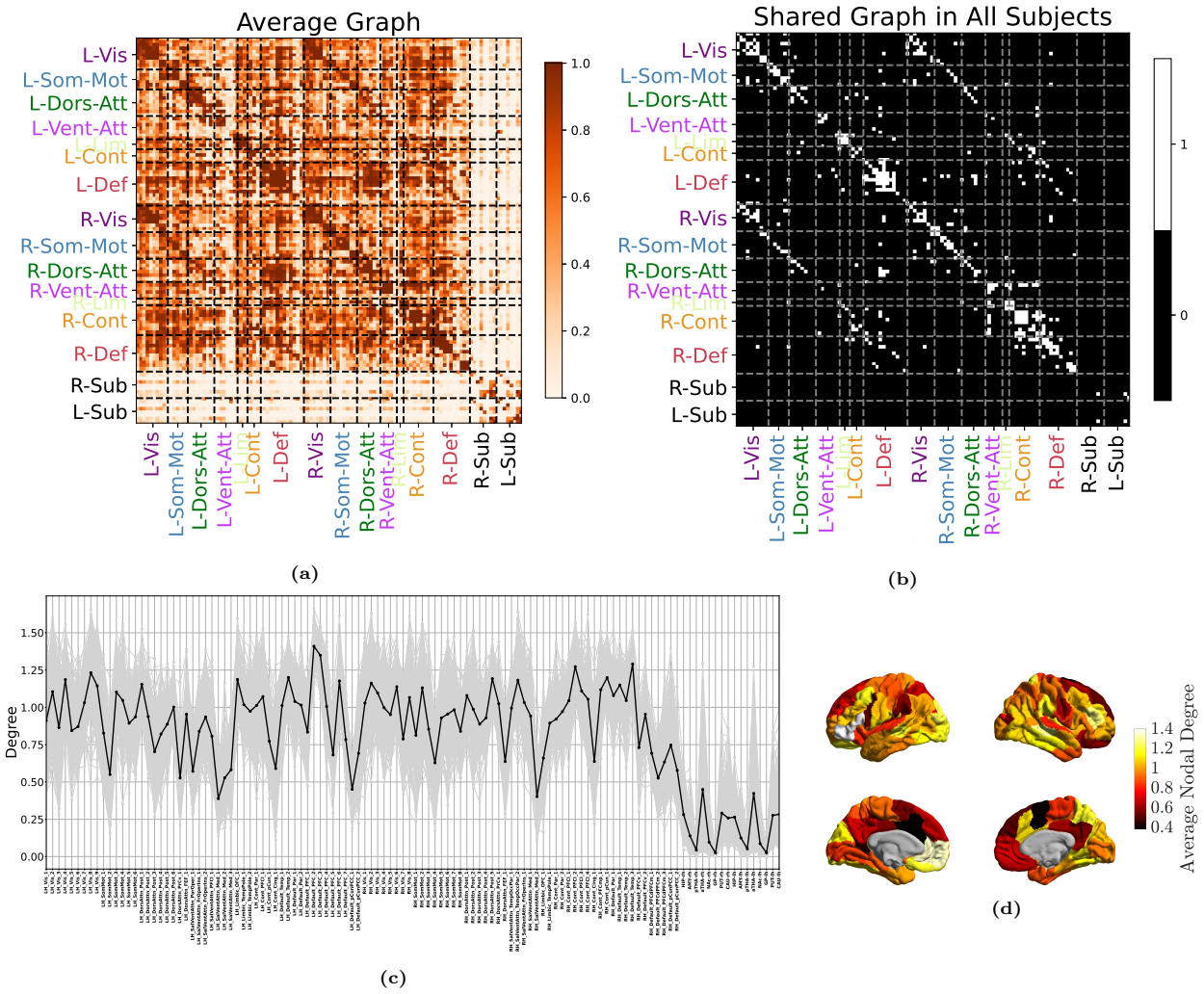


Figure 4: Learned causal graphs are highly consistent across subjects. (a) Average causal graph, computed as the mean of all the causal graphs of 200 subjects. The color of the (i, j) entry in this matrix shows how likely it is for node i to cause node j. A clear distinction can be seen in the causal links among cortical regions (colored labels) and subcortical ones (black labels). **(b)** The shared causal graph, containing only the edges that are present in all 200 subjects. This graph is dominated by diagonal elements (self- and within-subnetwork causation) and links among symmetrically located regions across the two hemispheres. **(c)** Distribution of nodal degree, computed separately for each node in the causal graph and each subject. Gray colors correspond to different subjects and the black line shows the average nodal degrees across subjects. **(d)** Average cortical nodal degree (black line in (c)) shown over the brain cortical surface.

Net resting-state causal effect flows from attention and default mode to sensorimotor networks. One of the main advantages of directed causal connectomes over undirected functional and structural connectomes is the former's ability to show the directed flow of causal effect between brain regions. In graphs learned by CaLLTiF, nodal causal flows (outflow minus inflow, see Methods) are also highly consistent across subjects (Figure 5c,5d), even though the two notions of centrality are generally dissociated across parcels (Figure 5b and Supplementary Figure 51). On average across all subjects, we observed particularly high causal flows (source-ness) in several regions of bilateral medial ventral attention networks, specific dorsal attention areas (ventral precentral, ventral frontal cortices, and frontal eye fields), and bilateral hippocampi, even though subcortical areas are much less connected to the rest of the network in general. In contrast, bilateral visual areas show the strongest negative causal flow (sink-ness) across all subjects. There is also notable variability among parcels within a subnetwork, such as the notable bilateral contrast between the strongly positive and weakly negative causal flows of frontal and posterior parts of the dorsal attention network, respectively.

To better assess the overall net causal effects between different functional networks, we computed an average "subnetwork graph" in which each node represents a functional cortical network and edges denote thresholded average directed connectivity from one functional network to another (see Methods for detailed computations). The result is illustrated in Figure 5a. Ventral attention and visual networks are clearly the strongest source and sink of causal flow, respectively. The dorsal attention and somatomotor networks are also a clear (though weaker) source and sink, respectively. The default mode network (DMN) is also a net source of causal flow, even though its outflows and inflows are nearly balanced. Similarly, the control and limbic networks have near-zero causal flows (near-balanced inflow and outflow). Several directed paths, however, can be seen from both attention and default mode networks to sensorimotor networks through the limbic and control networks. Therefore, in summary, causal graphs learned by CaLLTiF show the strongest net resting state causal effect to flow from the ventral and dorsal attention as well as the default mode networks, through control and limbic networks, towards sensorimotor networks, with the DMN serving both as a hub and a weak source of causal flow (cf. Discussion for a more detailed summary).

Casual graphs are strongly dominated by contemporaneous and lag-1 connections. Given that the final causal graph returned by CaLLTiF is a union over subgraphs at different lags (cf. Methods), we can go back and ask how much causal effects in each lag have contributed to the final graph. Figure 6a shows the percentage of edges in the final graph that exist *only* in one lag (including lag 0, or contemporaneous edges). Increasing the lag order resulted in significantly sparser single-lag subgraphs, which contributed less to the end result. In particular, approximately 70% of the end graphs came from lag 0 alone, a pattern that appears consistently across all subjects (Supplementary Figure 55). Even further, such contemporaneous edges are substantially stronger than edges from lags 1-3 (Figure 6b). This further confirms that the contemporaneous effects are particularly important for fMRI, where most neural dynamics occur at timescales shorter than 1 TR (typically shorter than 1-2 seconds). This is even the case in HCP data, with TR = 0.72s which is among the shortest TRs currently available in fMRI research. That being said, all lags had a non-zero (and significant by construction) contribution to the end graph in all subjects. Even lag 3 had a median of approximately 0.2% unique contributions to the final graph across subjects. We also found very small intersections among lags. This not only highlights the importance of considering multiple lags rather than just the first one or two but also demonstrates that it is incorrect to assume that if one region causes another, that causation will appear continuously across all lags. In summary, we found contemporaneous effects dominant in the final causal graphs of CaLLTiF, even though all lags had significantly non-zero and mostly unique contributions.

Causal connections are modulated by pairwise Euclidean distance. As one would expect from a network learned over a set of nodes embedded in physical space, the causal graphs learned by CaLLTiF are modulated in a number of ways by the Euclidean distance between pairs of nodes. First, we found degree similarity (correlation coefficient between nodal degrees of two parcels over all subjects) to decay statistically significantly, though weakly in effect size, with parcel distance (Pearson $r = -0.12, p = 10^{-43}$, 95% confidence interval $(-0.14, -0.1)$) as shown in Figure 7a (See Supplementary Figures 52 and 53 for separate maps of degree similarities and pairwise nodal distances). This relationship is stronger among intra-hemispheric parcels (Pearson $r = -0.27, p = 10^{-82}$, 95% confidence interval $(-0.29, -0.24)$) where connections are denser

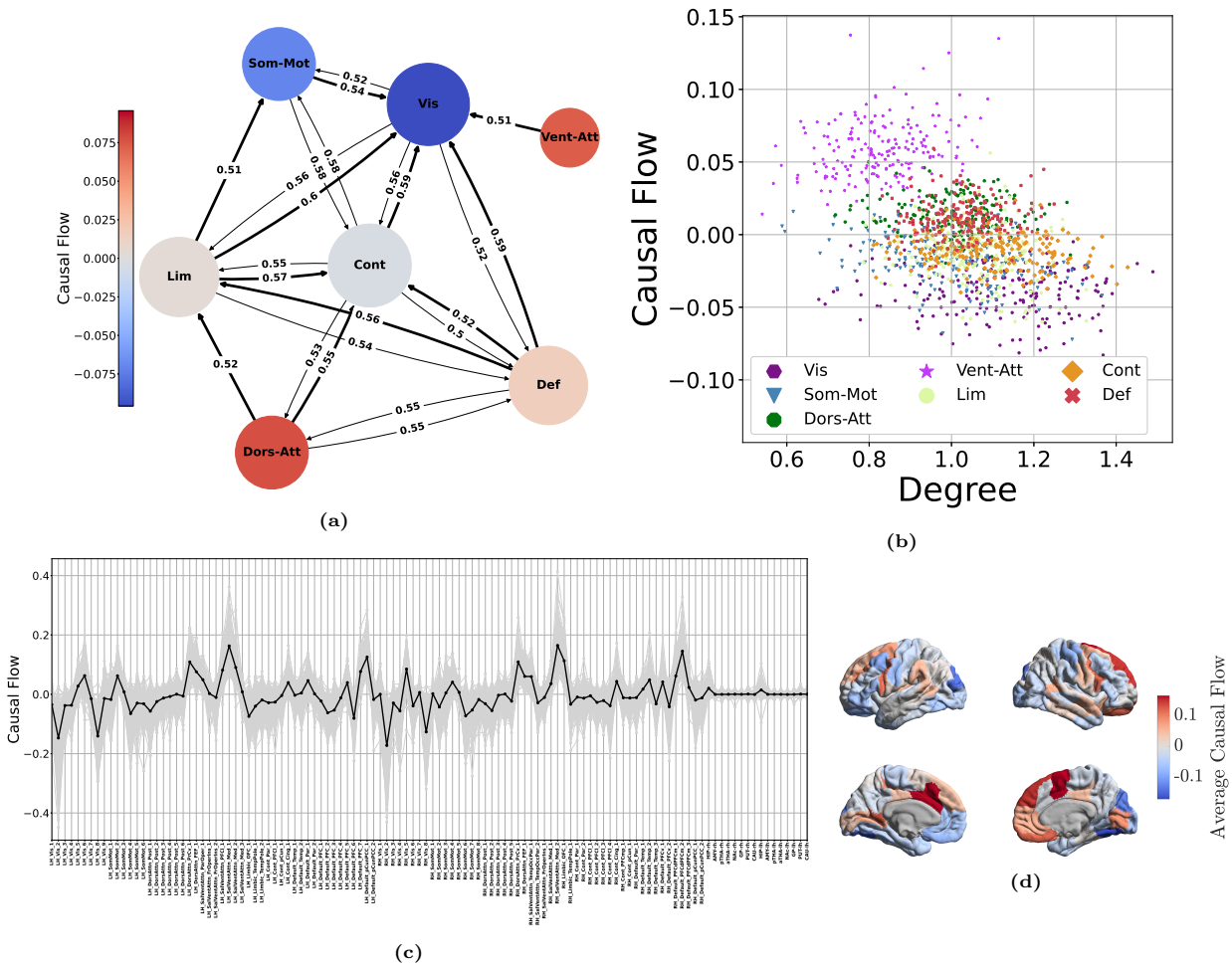


Figure 5: Patterns of causal flow across subjects. **(a)** The average subnetwork graph, computed as the mean of subnetwork graphs of all the subjects. In the subnetwork graph of each subject, the weight of an edge from subnetwork i to j is the number of nodes in subnetwork i that connect to nodes in subnetwork j, normalized by the number of all possible edges between these subnetworks. Edges with weights less than 0.5 are removed for better visualizations. To further ease the visual inspection of causal flows, if two networks are bidirectionally coupled we have shown the stronger edge with a thicker line (see Supplementary Figures 49, 50 for the weight matrix before thresholding and the corresponding nodal degrees and causal flows.) **(b)** The joint distributions of causal flow and degree for each "node" of the subnetwork graphs across all subjects. **(c)** Distribution of nodal causal flow, computed separately for each node in the causal graph and each subject. Gray colors correspond to different subjects and the black line shows the average nodal degrees across subjects. **(d)** Average cortical nodal causal flows (black line in (c)) shown over the brain cortical surface.

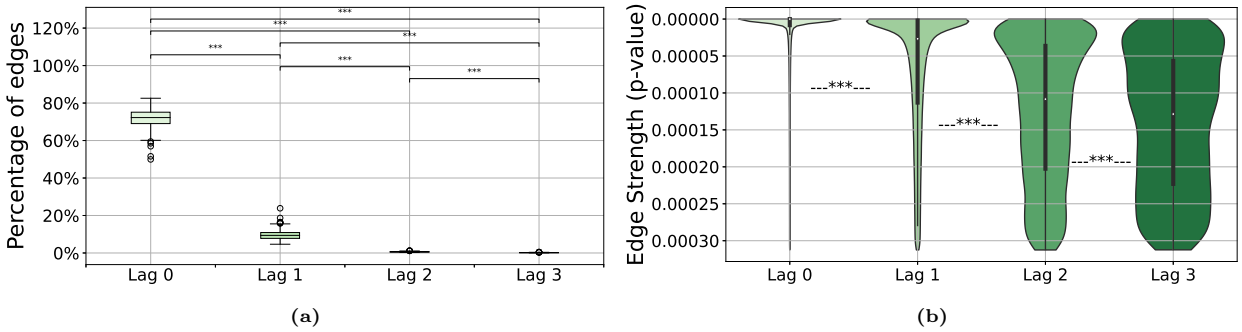


Figure 6: The contributions of each lag to the final causal graph in CaLLTiF. (a) For each lag, the box plot shows the distribution of the percentage of edges that come only from that lag across subjects. In all boxplots, the center line represents the median, the box spans the interquartile range (IQR), the whiskers extend up to 1.5 times the IQR from the box limits, and individual points beyond the whiskers indicate outliers.**(b)** The strength (statistical significance) of edges across lags. For each edge within the subgraph of each lag, we define its strength as the p-value of the partial correlation test that was used to conclude the presence of that edge (see Methods) even though all edges have a statistically significant p-value by definition, edges in larger lags are significantly closer to the threshold for significance than those in smaller lags. $^{***} = p < 0.001$, one-sided Wilcoxon rank-sum test.

and shorter-distance, compared to inter-hemispheric parcels (Pearson $r = -0.09, p = 10^{-5}$, 95% confidence interval $(-0.13, -0.05)$). Thus, in summary, nodes that are physically closer to each other also have more similar causal connections to the rest of the network, particularly if they belong to the same hemisphere.

The strength of CaLLTiF edges is also modulated by the Euclidean distance between edge endpoints, even though we observed that there are approximately as many long-distance edges as short ones (See Supplementary Figures 54). We define the strength of each edge in the final graph (union over lags) as the *minimum* p-value of respective partial correlation tests across all lags (cf. Methods). As seen from Figure 7b, the mean strength of causal edges (black solid line) initially increases with the Euclidean length of the edge until about 20mm and then decays with Euclidean edge length thereafter.

Finally, we found no major differences between the Euclidean distances of edges contributed by different lags. Given that causal effects take time to spread along axonal fibers throughout the brain, one might expect physically-closer pairs of nodes to be connected by lower-lag edges and more distant pairs of nodes to be connected by larger-lag edges. However, as seen in Supplementary Figure 56, this is not quite the case. Given the slow sampling of fMRI, even the most distant regions can causally affect each other in time scales shorter than 1 TR. Thus, the observation that the physical distance of pairs of nodes was not related to edge lag should not be taken as an indication that such relationships would – or would not – be observed when sampling with higher temporal precision.

Degree, but not casual flow, shows significant laterality and gender differences. We observed that nodal degrees were statistically significantly higher in the right hemisphere (Figure 8a, Cohen's d = 0.07 and $p = 10^{-48}$, one-sided Wilcoxon signed-rank test), even though no such laterality was found in nodal causal flows (Figure 8b, Cohen's d = 0.02 and $p = 0.23$, one-sided Wilcoxon signed-rank test). To understand which subnetworks might be playing a stronger role in the hemispheric asymmetry observed in the distribution of nodal degrees, Figure 8c shows the mean degrees of corresponding pairs of regions in the left and right hemispheres, color-coded by functional subnetworks (cf. Supplementary Figures 47 for separate plots per subnetwork). The ventral attention, dorsal attention, and executive control networks show clearly larger causal degrees in the right hemisphere, whereas the limbic network and DMN have larger causal degrees in the left hemisphere. A similar plot for causal flows (Figure 8d, Supplementary Figure 48) shows a lot more symmetry, except for the limbic network which shows exceptionally higher causal flows (i.e., source-ness) in the right compared to the left hemispheres. The DMN also shows some asymmetry in its causal flow, where right DMN nodes are mostly sources of causal flow whereas left DMN causal flows are more evenly distributed around zero. Thus, in summary, various functional subnetworks show laterality in degree distributions, culminating in a net increase in right vs. left nodal degrees. Causal flows, however, are mostly symmetric, except for the limbic network which shows a strong flow from the right to the left hemisphere.

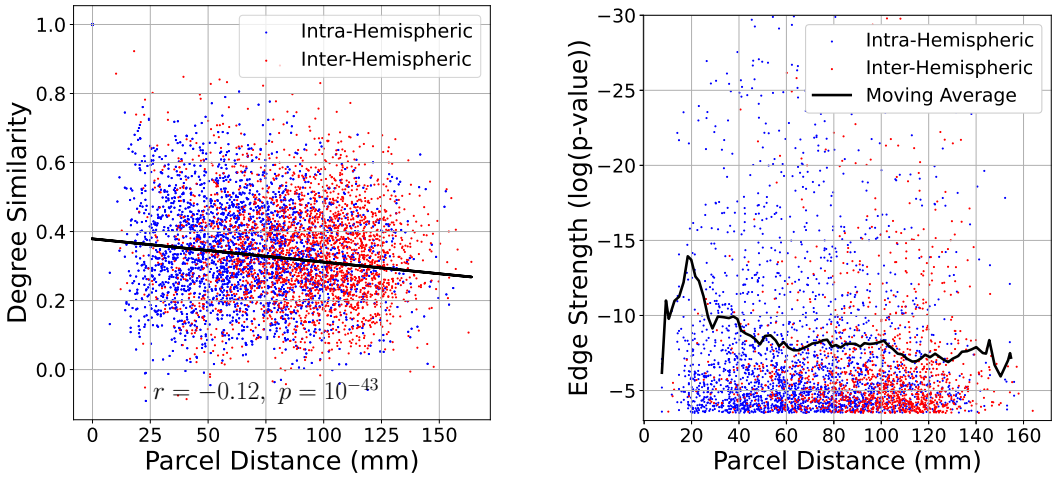


Figure 7: Effect of Euclidean distance on edge attributes. (a) Degree similarity (correlation coefficient between nodal degrees of two parcels over all subjects) as a function of the Euclidean distance between the parcels. Parcel pairs in the same hemisphere (intra-hemispheric) and parcels in two different hemispheres (inter-hemispheric) are shown in blue and red, respectively. Degree similarity decays statistically significantly with parcel distance (Pearson $r = -0.12$, $p = 10^{-43}$, 95% confidence interval $(-0.14, -0.1)$), much more so among intra-hemispheric parcels (Pearson $r = -0.27$, $p = 10^{-82}$, 95% confidence interval $(-0.29, -0.24)$) than inter-hemispheric ones (Pearson $r = -0.09$, $p = 10^{-5}$, 95% confidence interval $(-0.13, -0.05)$). **(b)** Edge strength (as in Figure 6b) as a function of Euclidean distance between edge endpoints (note the inverted vertical axis). The solid line shows the corresponding moving average of $\log(p)$ with 10mm window size and 8mm window overlap. The upper limit of the vertical axis is limited to -30 for better visualization.

Similarly, degree, but not casual flow, shows a small but statistically significant difference between men and women. In causal connectomes learned by CaLLTiF, we found nodal degrees to be statistically significantly higher in women compared to men (Figure 8e, Cohen's d > 0.05, $p < 10^{-5}$, one-sided Wilcoxon rank-sum test). Nodal causal flows, on the other hand, where statistically indistinguishable between men and women (Figure 8f, Cohen's d $= -3.77 \times 10^{-18}$, $p = 0.81$, one-sided Wilcoxon rank-sum test). These result demonstrate that nodal degrees in causal connectomes are generally more heterogeneous and sensitive while causal flows tend to be more homogeneous and stereotyped across individuals and hemispheres. Further research is needed to pinpoint the root causes of these differences (and lacks thereof) as well as potential implications of them in health and disease.

Discussion

In this study, we investigated the problem of whole-brain causal discovery from fMRI. We first comprehensively compared existing causal discovery techniques suitable for whole-brain fMRI by examining both theoretical properties and numerical outcomes on simulated fMRI. To address the limitations of existing algorithms, we proposed CaLLTiF which improves upon the state of the art in several directions including learning contemporaneous edges and cycles, type I error control, and scalability. A core aspect of CaLLTiF is its treatment of contemporaneous effects. Our results with the HCP data (Figures 6a and 6b) confirmed the importance of being able to reveal such "contemporaneous" effects, where these effects accounted for the majority and strongest of network edges. Further, the distributions of edges with different Euclidean distances at each lag (Supplementary Figure 56) demonstrates how broadly neural signals can propagate across the brain in one TR interval, even with the relatively fast sampling (TR = 0.72s) in the HCP dataset.

Furthermore, in interpreting CaLLTiF's outputs, it is important to note its by-design conservative method of correction for temporal multiple comparisons. In simulated Full Macaque data where the ground truth is known, we found Alpha Level = 0.01 to maximize the F1 score, while CaLLTiF's correction for temporal multiple comparisons would have suggested $0.01/32 = 0.0003$ (cf. Equation (4)) and thus would have obtained sparser graphs. Similarly, we obtained causal graphs from human fMRI that are about 30-55% dense across

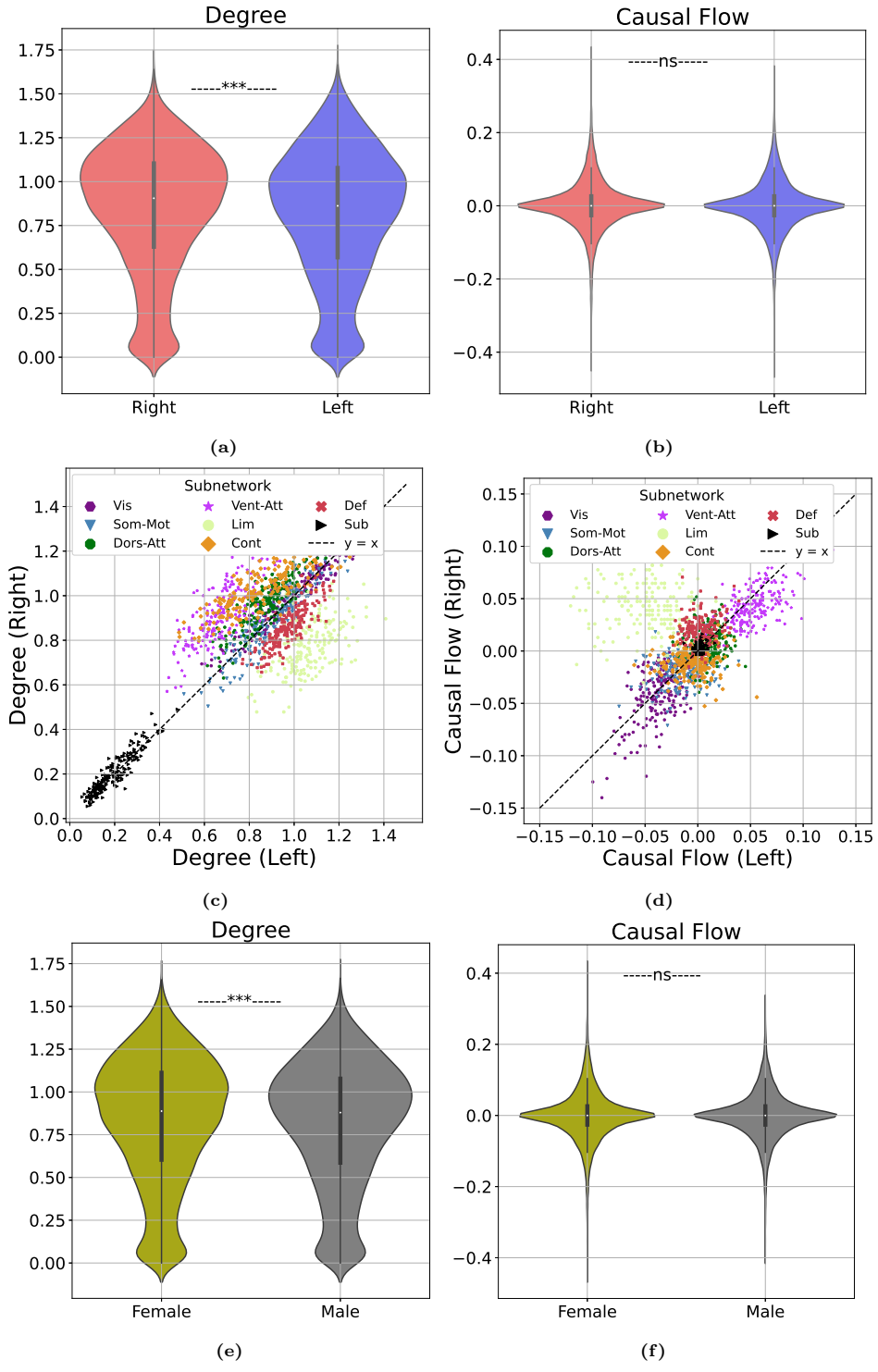


Figure 8: Hemispheric laterality and gender differences in causal connectomes. (a) Distributions of nodal degree in the right and left hemispheres, combined across all subjects. (b) Similar to (a) but for causal flows. (c) Nodal degrees, averaged across subjects and color-coded by functional subnetwork, for pairs of corresponding parcels within the right and left hemispheres. To properly pair nodes across the two hemispheres, degrees of all the parcels with the same label in the Schaefer/Tian atlas were averaged and then paired. Deviations from the dashed y = x line indicate laterality. (d) Similar to (c) but for causal flows. (e) Distributions of nodal degrees in female and male subjects, combined across all brain regions. (f) Similar to (e), but for causal flows.

all subjects and 40-60% dense among cortical nodes (Supplementary Figure 41a). These are generally consistent with, but sparser than, the near-66% cortical density found using tract-tracing results in non-human primates (Markov et al., 2014). In other words, graphs returned by CaLLTiF are likely to have higher precision but lower recall than what would maximize the F1 score. This conservativeness is by design and desirable *in the absence of a ground-truth causal graph*, ensuring that we have strong confidence in all discovered edges (at least 99% confidence in every detected edge in our experiments with human fMRI). Nevertheless, CaLLTiF's level of conservativeness can also be tuned as needed by tuning its pre-correction significance threshold (q in Equation (4)).

An unexpected finding of our study was the higher accuracy of causal discovery when conditioning pairwise independence tests (see Equation (1)) on all other nodes in the network, as done in CaLLTiF, compared to using a more restricted parent set found by PCMCI (cf. Supplementary Figure 16). The approach taken by PCMCI increases statistical power (cf. the trend of optimal 'Alpha Level' values in Supplementary Figure 16b), but can significantly increase type I error in the presence of contemporaneous effects. Even further, we found that even using the (lagged) ground-truth parent sets for each node leads to a lower F1 score compared to using complete conditioning sets (Supplementary Figure 38-40). This is likely because CaLLTiF's conditioning on the *past* of all variables serves as a proxy for the missing contemporaneous parents that should have been conditioned on. On the other hand, one may wonder if this issue could have been better resolved by conditioning on contemporaneous variables themselves. However, this can result in spurious statistical dependence if conditioning on all contemporaneous variables (consider, e.g., testing $X_i \not\perp\!\!\!\perp X_j | X_k$ with the ground-truth causal graph $X_i \to X_k \leftarrow X_j$). For a detailed discussion on this see Supplementary Note 3.

Causal connectivity during resting state. When applying CaLLTiF to resting state human fMRI, we found the strongest causal effect to flow from attention to sensorimotor networks. The strongest sources and sinks were the ventral attention and visual networks, followed by dorsal attention and somatomotor networks, respectively. Despite the lack of "ground-truth" connectivity as in simulated data, we can still evaluate these findings based on their agreement with prior findings on the roles of resting state networks. The dorsal attention network (involving regions in the intraparietal and superior frontal cortex) is commonly believed to handle top-down selection processes and is only modulated by stimulus detection, while the ventral attention network (including areas such as the temporoparietal and inferior frontal cortex) specializes in detecting behaviorally relevant stimuli, particularly those that are salient or unexpected, and directing attention to them (Corbetta and Shulman, 2002). These networks exhibit internally correlated activity patterns (particularly during the task) (Corbetta et al., 2008) and their flexible interaction facilitates dynamic attentional control aligned with both top-down goals and bottom-up sensory inputs (Vossel et al., 2014). Nonetheless, the ventral attention network is more strongly involved in the "reorientation" of attention, namely, interrupting one thought process and orienting attention towards something salient, while the dorsal attention network is more strongly implicated in focused and guided attentional tasks such as visual search under high short-term memory load (Vossel et al., 2014).

Arguably, resting state activity is more aligned with the former (salience-based reorientation) than the latter (goal-driven focused attention). Despite a lack of sensory salience, attention is frequently reoriented during periods of rest by endogenously-salient thoughts and memories. Intermittent periods of focused attention can also arise, particularly given the long durations of each resting-state scan ($\sim$ 15min). Our findings thus confirm and corroborate the existing hypotheses about the roles of attention networks and how they jointly but unevenly drive brain dynamics during rest. Furthermore, due to the lack of meaningful sensory (particularly visual) input during a resting state scan, sensory areas receive more top-down influence from higher-order cortices than they provide bottom-up information to them. As such, the sink-ness of sensorimotor areas in graphs learned by CaLLTiF is arguably more consistent with the nature of resting state activity compared to a contrasting, sensory-driven flow found, e.g., in (Rawls et al., 2022). Finally, we found the DMN to be both a hub and a weak source of causal flow, which is consistent with its well-known role in resting state dynamics (Andrews-Hanna et al., 2014; Greicius et al., 2003; Raichle et al., 2001).

Resting-state causal graphs learned by CaLLTiF are also notably aligned with the literature in terms of the laterality of different functional subnetworks (Figure 8c and Supplementary Figure 47). Several studies have found the ventral attention system to be predominantly lateralized to the right hemisphere (Corbetta et al., 2008; Corbetta and Shulman, 2002; Mengotti et al., 2020; Vossel et al., 2014). Similarly, the degree

distribution of ventral attention nodes in graphs learned by CaLLTiF is strongly right-lateralized. We found the dorsal attention network to also be right-lateralized, but not as strongly as the ventral network. Similarly, The dorsal attention network is found by prior studies to be organized mostly bilaterally, except for specific regions (Intraparietal sulcus and frontal eye field) in the right dorsal network that show stronger involvement in the attentional control of both hemispheres compared to their left counterparts (Mengotti et al., 2020). Also similar is alignment in the lateralization of the default mode network, where both CaLLTiF several past studies have found it to be left lateralized (Agcaoglu et al., 2015; Banks et al., 2018; Nielsen et al., 2013; Swanson et al., 2011). Sensorimotor cortices, on the other hand, were found to be highly symmetric and not particularly lateralized to either hemisphere in causal graphs learned by CaLLTiF, a finding that is also consistent with the generally symmetric involvement of primary sensory and motor areas in contralateral processing (Agcaoglu et al., 2015). Finally, research on the laterality of the control and limbic networks is ongoing and, to the best of our knowledge, inconclusive (see, e.g., (Morton, 2020)). In graphs learned by CaLLTiF, however, we observe strong left lateralization of the limbic and right lateralization of the control networks, respectively. Thus, in summary, we observe clear laterality in all but sensorimotor cortical networks which either corroborate the existing literature or clarify previously inconclusive observations.

Limitations. The present study has a number of limitations. The TR value of 720ms in the HCP data limits the precision of causal discovery. As we saw from Supplementary Figure 56, edges of all lengths are observed even at lag 0. This indicates the possibility that some of the edges discovered by CaLLTiF may be polysynaptic paths but resemble a direct monosynaptic connection at low temporal resolution. Moreover, similar to most constraint-based methods, the causal graphs returned by CaLLTiF are not tied to a generative dynamical model (as is the case with VARLiNGAM, DYNOTEARS, DCM, etc). If such generative models are needed, VAR models based on CaLLTiF's extended time-lagged graph constitute a natural choice, but further research is needed to compare the dynamic predictive accuracy of such models against potential alternatives (Ljung, 1999).

Conclusions. Overall, this study demonstrates the interplay between the theoretical challenges of causal discovery and the practical limitations of fMRI, and the design of an algorithmic solution that can bridge this gap. This work motivates several follow-up studies, including the application of the proposed CaLLTiF method to task fMRI and comparing its outcomes against structural connectivity.

Material and methods

Simulated fMRI Data

When comparing different causal discovery algorithms or different hyperparameters of the same algorithm, we used several benchmarks of simulated fMRI data with known ground truth connectivity from (Sanchez-Romero et al., 2019). In general, this dataset included two groups of networks, one consisting of 9 simple small-scale synthetic graphs and one consisting of two graphs extracted from the macaque connectome. From the latter group, we only used the smallest ('Small-Degree Macaque') and the largest ('Full Macaque').

The details of generating BOLD signals from each graph are detailed in (Sanchez-Romero et al., 2019). In brief, the same simulation procedure was used for simple and macaque-based graphs, where the authors used the model proposed in (Smith et al., 2011) which is itself based on the DCM architecture of (Friston et al., 2003). Underlying neural dynamics are simulated using the linear differential equation $dz/dt = \sigma Az + Cu$, where A denotes the ground-truth connectivity. To simulate resting-state data, the u input was modeled using a Poisson process for each of the regions ($C = I$). The neuronal signals z were then passed through the Balloon-Windkessel model (Buxton et al., 1998; Smith et al., 2011) to obtain simulated BOLD data.

Resting-State fMRI from the Human Connectome Project

For the real fMRI analysis, we used ICA-FIX resting-state data from the Human Connectome Project S1200 release (Barch, 2017; Burgess et al., 2016; Essen et al., 2013). Resting-state fMRI images were collected with the following parameters: TR = 720 ms, TE = 33.1 ms, flip angle = 52 deg, FOV = 208x108 mm, matrix

= 104x90, slice thickness = 2.0 mm, number of slices = 72 (2.0 mm isotropic), multi-factor band = 8, and echo spacing = 0.58 ms. Brains were normalized to fslr32k via the MSM-AII registration and the global signal was removed. We removed subjects from further analysis if any of their four resting state scans had excessively large head motion, defined by having frames greater than 0.2 mm frame-wise displacement or a derivative root mean square (DVARS) above 75. Also, subjects listed in (Elam, 2020) under "3T Functional Preprocessing Error of all 3T RL fMRI runs in 25 Subjects" or "Subjects without Field Maps for Structural scans" were removed. Among the remaining 700 subjects, the 200 with the smallest head motion (DVARS) were selected for analysis. For all subjects, we parcellated the brain into 100 cortical regions (Schaefer 100x7 atlas (Schaefer et al., 2018)) and 16 subcortical ones (Melbourne Scale I atlas (Tian et al., 2020)). The Human Connectome Project experiments were carried out by the WU-Minn consortium and its adherence to ethical standards was approved by the by the Internal Review Board of the respective institutions. Explicit informed consent was acquired from all participants involved in the study (Essen et al., 2013).

Causal discovery methods

One aim of causal inference is to construct a causal graph based on observational data. The relationship between a probability distribution and its depiction as a graph plays a significant role in this process. Nevertheless, it is not always feasible to deduce a causal graph solely from observational data. Further assumptions are therefore required. Here, we briefly summarize the main assumptions and principles underlying the list of causal discovery methods studied in this work (cf. Table 1).

PCMCI. PCMCI was proposed in (Runge et al., 2019) as a constraint-based causal discovery method designed to work with time-series data. The algorithm is composed of two main steps. In the first step, the algorithm selects relevant variables using a variant of the skeleton discovery part of the PC algorithm (Spirtes and Glymour, 1991). This step removes irrelevant variables for conditioning and therefore increases statistical power. In the second step, the algorithm uses the momentary conditional independence (MCI) test, which measures the independence of two variables conditioned on the set of their parents identified in step 1. The MCI test helps to reduce the false positive rate, even when the data is highly correlated. PCMCI assumes that the data is stationary, has time-lagged dependencies, and has causal sufficiency. Even when the stationarity assumption is violated, PCMCI was shown to perform better than Lasso regression or the PC algorithm (Runge et al., 2019). However, PCMCI is considered not suitable for highly predictable (almost deterministic) systems with little new information at each time step (Runge et al., 2019). The Python implementation of PCMCI is available in the Tigramite package at `https://github.com/jakobrunge/tigramite`.

As noted earlier, PCMCI only returns $\circ\!\!-\!\!\circ$ edges among contemporaneous variables. While this allows PCMCI to relax the common DAG assumption and allow for cycles, it results in a mixed summary graph, where multiple types of edges ($\leftarrow$, $\rightarrow$, and/or $\circ\!\!-\!\!\circ$) can exist between two nodes. In contrast, we require all algorithms to output a directed graph. Therefore, when reporting F1 scores for PCMCI, we only include directed edges coming from lagged relationships and exclude the contemporaneous $\circ\!\!-\!\!\circ$ edges. The only exception is what we call 'Mixed PCMCI' (See Supplementary Figures 27- 29), where the contemporaneous $\circ\!\!-\!\!\circ$ edges are also included in the computation of *adjacency* F1 scores.

PCMCI$^+$. PCMCI$^+$ is an extension of the PCMCI method which incorporates directed contemporaneous links in addition to the lagged ones (Runge, 2020). The approach revolves around two key concepts. First, it divides the skeleton edge removal phase into separate lagged and contemporaneous conditioning phases, thereby reducing the number of conditional independence tests required. Second, it incorporates the idea of momentary conditional independence (MCI) tests from PCMCI (Runge et al., 2019) specifically in the contemporaneous conditioning phase. PCMCI$^+$ also outputs a time-series graph with different types of contemporaneous edges, including directed edges ($\rightarrow$ and $\leftarrow$), unoriented edges ($\circ\!\!-\!\!\circ$), and conflicting edges ($\times - \times$). Consistent with our requirement of a regular digraph at the end, we disregarded the unoriented and conflicting edges and retained only the directed ones. Similar to most other causal discovery algorithms, PCMCI$^+$ does not permit cycles in the contemporaneous links, which could potentially account for its relatively underwhelming performance over fMRI data. The Python implementation of PCMCI+ is also available in the Tigramite package `https://github.com/jakobrunge/tigramite`.

VARLiNGAM. VARLiNGAM is a causal discovery method that combines non-Gaussian instantaneous models with autoregressive models. This method, proposed in (Hyvärinen et al., 2010), builds on the fact that in the absence of unobserved confounders, linear non-Gaussian models can be identified without prior knowledge of the network structure. VARLiNGAM is capable of estimating both contemporaneous and lagged causal effects in models that belong to the class of structural vector autoregressive (SVAR) models and provides ways to assess the significance of the estimated causal relations. These models are a combination of structural equation models (SEM) and vector autoregressive (VAR) models. In addition, VARLiNGAM emphasizes the importance of considering contemporaneous influences, as neglecting them can lead to misleading interpretations of causal effects. Nevertheless, VARLiNGAM does not permit cycles in the contemporaneous links either, which could potentially account for its relatively poor performance over brain fMRI data with many feedback loops. The VARLiNGAM method is available from `https://github.com/cdt15/lingam` and a tutorial can be found at `https://lingam.readthedocs.io/en/latest/tutorial/var.html`.

DYNOTEARS. Dynamic NOTEARS (DYNOTEARS) method, proposed in (Pamfil et al., 2020), is a score-based method designed to discover causal relationships in dynamic data. It simultaneously estimates relationships between variables within a time slice and across different time slices by minimizing a penalized loss function while ensuring that the resulting directed graph is acyclic (including acyclicity of contemporaneous connections). The goal is to identify the best set of conditional dependencies that are consistent with the observed data. DYNOTEARS builds on the original NOTEARS method proposed in (Zheng et al., 2018), which uses algebraic properties to characterize acyclicity in directed graphs for static data. Python implementations are available from the CausalNex library (`https://github.com/quantumblacklabs/causalnex`) as well as `https://github.com/ckassaad/causal_discovery_for_time_series`.

DGlearn. DGlearn is a score-based method for discovering causal relationships from observational data. Importantly, it is one of few algorithms that can learn cyclic structures from cross-sectional data. The method, introduced in (Ghassami et al., 2020), is based on a novel characterization of equivalence for potentially cyclic linear Gaussian directed graphical models. Two structures are considered equivalent if they can generate the same set of data distributions. DGlearn utilizes a greedy graph modification algorithm to return a graph within the equivalence class of the original data-generating structure. The Python implementation of DGlearn is available at `https://github.com/syanga/dglearn`.

FASK. The Fast Adjacency Skewness (FASK) method, proposed in (Sanchez-Romero et al., 2019), is a hybrid method for causal discovery from cross-sectional data, combining constraint-based and noise-based elements. It leverages (and needs) non-Gaussianity in the data and allows for cycles. This algorithm is composed of two main steps. The first step, called FAS-Stable, outputs an undirected graph G_0 by iteratively performing conditional independence tests under the increasing size of the conditioning set and using the Bayesian information criterion (BIC) to compare the conditioning sets. In the second step, assuming i.i.d. non-Gaussian data, each of the $X - Y$ adjacencies in G_0 are oriented as a 2-cycle ($\leftrightarrows$) if the difference between $corr(X, Y)$ and $corr(X, Y|X > 0)$, and $corr(X, Y)$ and $corr(X, Y|Y > 0)$, are both significantly nonzero, and as a unidirectional edge otherwise. The pseudo-code for FASK can be found in Supporting Information A of (Sanchez-Romero et al., 2019) and Java source code for it is available at `http://github.com/cmu-phil/tetrad`.

MVGC. In (Granger, 1969), Granger introduced a statistical version of Hume's regularity theory, stating that X_p Granger-causes X_q, if past values of X_p provide unique, statistically significant information about future values of X_q (Assaad et al., 2022). While this allows for optimal forecasting of an effect and has been extended to multivariate systems (Barnett and Seth, 2014), MVGC cannot account for contemporaneous effects and the presence of unobserved confounders can result in spurious edges. Python implementation of MVGC is available at `https://github.com/ckassaad/causal_discovery_for_time_series`.

NTS-NOTEARS. NTS-NOTEARS is a nonlinear causal discovery method designed for time-series data (Sun et al., 2021). It employs 1-D convolutional neural networks to capture various types of relationships,

including linear, nonlinear, lagged, and contemporaneous connections among variables. The method ensures that the resulting causal structure forms a directed acyclic graph. It builds upon the NOTEARS approach (Zheng et al., 2018), and is similarly based on continuous optimization. Similar to other algorithms above, it assumes the presence of no hidden confounding factors and stationarity of the data-generating process. In our analysis, we compare NTS-NOTEARS as a state-of-the-art nonlinear method against the aforementioned linear algorithms in synthetic fMRI (cf. Supplementary Figure 36). A Python implementation of NTS-NOTEARS is available at `https://github.com/xiangyu-sun-789/NTS-NOTEARS`

CaLLTiF (proposed method). The proposed CaLLTiF method builds upon PCMCI (Runge et al., 2019) but, instead of using a PC-type approach in the first step to estimate the set of parents for lagged variables, it starts from a complete conditioning set including all lagged variables. This choice dramatically decreases computational cost, but surprisingly, it is also optimal, as shown in Supplementary Figure 16, because as mentioned in the discussion section, the approach of PCMCI discards contemporaneous effects. Using a complete conditioning set, CaLLTiF then performs Momentary Conditional Independence (MCI) partial correlation tests between all pairs of variables. Specifically, for any pair $X_i(t-\tau), X_j(t)$ with $i, j \in 1, \ldots, N$ and time delays $\tau \in 0, 1, \ldots, \tau_{max}$, a causal link is established ($X_i(t-\tau) \to X_j(t)$ if $\tau > 0$ and $X_i(t) \circ\!\!-\!\!\circ X_j(t)$ if $\tau = 0$), if and only if:

$$\text{Cond. Ind. Test} : X_i(t-\tau) \not\perp\!\!\!\perp X_j(t) \mid \{X_k(t-s) : k = 1, \ldots, N, s = 1, \ldots, \tau_{\max}\} \setminus \{X_i(t-\tau)\}. \quad (1)$$

Note that, despite being complete, the conditioning sets only include variables from *prior* time lags. As noted earlier, to test a conditional independence of the form $X \not\perp\!\!\!\perp Y|Z$, we compute the partial correlation coefficient $\rho(X, Y|Z)$ between X and Y conditioned on the set of variables in Z and the corresponding p-value for the null hypothesis that $\rho(X, Y|Z) = 0$. An edge is placed between $X_j(t)$ and $X_i(t-\tau)$ if this p-value is less than the hyperparameter 'Alpha Level'. The value of this threshold was selected optimally in simulated fMRI and using temporal correction for multiple comparisons (see below) in real data. Finally, for contemporaneous pairs ($\tau = 0$), each $\circ\!\!-\!\!\circ$ edge is replaced with $\leftrightarrows$ if there are no other edges between those two variables at other lags, and is replaced with a directed edge or a $\leftrightarrows$ based on the lagged direction(s) otherwise. For a more detailed summary of the partial correlation-based edge discovery in CaLLTiF, see Supplementary Note 2.

Finally, it is imperative to acknowledge the possibility that some of the directed edges detected by our methodology do not possess a strictly causal connotation. As previously indicated, the orientation method relies on the widely accepted premise that bidirectional connections hold notably greater prevalence than unidirectional links. Thus, we believe that the presented approach shall yield a proximate representation of the true causal graph, concurrently accommodating cyclic structures and circumventing computational barriers.

Construction of summary causal graphs from causal graphs over lagged variables

Causal discovery algorithms designed for time series data often return a causal graph among the lagged variables

$$\begin{aligned} &X_1(t-\tau_{\max}), \ldots, X_1(t-1), X_1(t) \\ &X_2(t-\tau_{\max}), \ldots, X_2(t-1), X_2(t) \\ &\vdots \\ &X_n(t-\tau_{\max}), \ldots, X_n(t-1), X_n(t) \end{aligned} \quad (2)$$

From this, we extract a final *summary* graph among the variables $X_1, \ldots, X_n$ by placing an edge from X_i to X_j if there exists any $\tau \geq 0$ for which there is an edge from $X_i(t-\tau)$ to $X_j(t)$. This is equivalent to an OR operation among binary edges (as opposed, e.g., to a majority vote) and must be taken into account when interpreting the obtained summary graphs.

Correction for multiple comparisons across lags in CaLLTiF

As noted above, CaLLTiF places an edge from X_i to X_j in its summary graph if there exists at least one $\tau \geq 0$ for which there is an edge from $X_i(t-\tau)$ to $X_j(t)$. Therefore, the decision to place an edge from X_i to X_j depends on the outcomes of $\tau_{\max}+1$ statistical tests, and to maintain a desired bound on the probability of type I error for each edge in the *summary* graph, we need to account for multiple comparisons across lags.

Formally, for each edge $X_i \to X_j$ in the final graph, the null hypothesis (i.e., lack of a direct causal effect from X_i to X_j) can be formulated as

$$H_0 = H_{0,0} \cap H_{0,1} \cap \cdots \cap H_{0,\tau_{\max}} = \bigcap_{\tau=0}^{\tau_{\max}} H_{0,\tau},$$
$$H_{0,\tau} = \{X_i(t-\tau) \not\to X_j(t)\}.$$

Let p_τ denote the p-value of the partial correlation test between $X_i(t-\tau)$ and $X_j(t)$ and q denote the significance threshold for each partial correlation test. Then, the probability of type I error is

$$\text{P(Type I Error)} = P\Big(\bigcup_{\tau=0}^{\tau_{\max}} \{p_\tau < q\} \;\Big|\; \bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big)$$

Note that this is different from the family-wise error rate (FWER, bounded by the Bonferroni method and its extensions) or the false discovery rate (FDR). In particular, this is different from FWER in that only one decision is made and the probability of error is computed for that single decision only. So, for instance, if in reality any subset (even one) of $\{H_{0,\tau}\}$ is false and the algorithm rejects any subset (even all) of $\{H_{0,\tau}\}$, there is no type I error, as an edge exists from X_i to X_j both in the data-generating process and in the final summary graph.

The type I error can then be bounded as

$$\begin{aligned} P(\text{Type I Error}) &= P\Big(\bigcup_{\tau=0}^{\tau_{\max}} \{p_\tau < q\} \;\Big|\; \bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big) \leq \sum_{\tau=0}^{\tau_{max}} P\Big(\{p_\tau < q\} \;\Big|\; \bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big) \\ &= \sum_{\tau=0}^{\tau_{max}} \frac{P\Big(\{p_\tau < q\} \;\cap\; \bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big)}{P\big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\big)} \leq \sum_{\tau=0}^{\tau_{max}} \frac{P\Big(\{p_\tau < q\} \;\cap\; H_{0,\tau}\Big)}{P\big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\big)} \\ &= \sum_{\tau=0}^{\tau_{max}} \frac{P(H_{0,\tau})}{P\big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\big)} \frac{P\Big(\{p_\tau < q\} \;\cap\; H_{0,\tau}\Big)}{P(H_{0,\tau})} = q \sum_{\tau=0}^{\tau_{max}} \frac{P(H_{0,\tau})}{P\big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\big)} \end{aligned} \tag{3}$$

The last expression has no dependence on the data and depends only on the prior distribution we consider on graphs. Assuming a uniform prior, $P(H_{0,\tau}) = 1/2$. Further,

$$P\Big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big) = P(H_{0,0}|H_{0,1} \cap \cdots \cap H_{0,\tau_{\max}}) \cdots P(H_{0,\tau_{\max}-1}|H_{0,\tau_{\max}}) \cdot P(H_{0,\tau_{\max}})$$

We assume a prior where knowledge of the lack of an edge from X_i to X_j at one lag either increases the probability of lack of an edge between them at other lags or, at least, does not decrease it (independence across lags). Then,

$$P\Big(\bigcap_{\sigma=0}^{\tau_{\max}} H_{0,\sigma}\Big) \geq P(H_{0,0}) \cdots P(H_{0,\tau_{\max}-1}) \cdot P(H_{0,\tau_{\max}}) = \Big(\frac{1}{2}\Big)^{\tau_{\max}+1}$$

Putting everything together, we get

$$P(\text{Type I Error}) \leq q \frac{(\tau_{\max}+1)/2}{(1/2)^{\tau_{\max}+1}} = q(\tau_{\max}+1)2^{\tau_{\max}}.$$

Note, for analogy, that the correction factor $(\tau_{\max}+1)2^{\tau_{\max}}$ takes place of the factor $(\tau_{\max}+1)$ in a corresponding Bonferroni correction. To have $P(\text{Type I Error})$ less than a prescribed threshold α, we choose

$$q = \frac{\alpha}{(\tau_{\max}+1)2^{\tau_{\max}}} \tag{4}$$

In our experiments with the HCP data, we have $\tau_{max} = 3$ and $\alpha = 0.01$, giving a per-lag significance threshold of approximately 0.0003. This is notably smaller than the Alpha Level values that maximized F1 scores in simulated Full Macaque data (0.1 for adjacency F1 score and 0.01 for F1 score), and is due to the conservative nature of this correction for temporal multiple comparisons.

Computing Functional Graphs

In order to calculate the functional graphs for each subject, we consolidated the data from the four sessions of each subject in the HCP and computed the pairwise correlations among all pairs of parcels. To form a binary functional graph, we placed an edge between any two parcels displaying a statistically significant correlation coefficient ($p < 0.01$, t-test for Pearson correlation coefficient).

Hyperparameter Selection

All the methods we described in Table 1 have at least one main hyperparameter that significantly affects the end result, particularly in terms of edge density. These include 'PC Alpha' and 'Alpha Level' for PCMCI, 'PC Alpha' for PCMCI$^+$, 'Alpha' for VARLINGAM, DYNOTEARS, and MVGC, and FASK, and 'BIC Coefficient' for DGlearn. These hyperparameters were swept over (simultaneously for PCMCI) using the simulated data and selected such that the F1 score with the ground truth graph is maximized in each case. This process was repeated for all algorithms and all experiments (simple graphs, Small-Degree Macaque, Full Macaque). Performance metrics such as Recall, Precision, and F1 scores of each method for a range of their hyperparameters are shown in Supplementary Figures 6-12 for the simulated Simple Network graphs, in Supplementary Figures 20-24 for the simulated Small-Degree Macaque data, and in Supplementary Figures 30-36 for the simulated Full Macaque data.

Time-series algorithms (PCMCI, PCMCI$^+$, VARLiNGAM, DYNOTEARS) also have a hyperparameter controlling the number of lags used for causal discovery. Based on our prior work (Nozari et al., 2023), we set this value to 3 for the HCP data (TR = 0.72s), and confirmed its sufficiency based on the contributions of higher-order lags (Figure 6a). For the simulated data, (TR = 1.2s), we used a maximum lag of 2 to match its slower sampling.

Computing F1 Scores, Degrees, and Causal Flows

In our experiments using simulated fMRI data, access to ground truth graphs allows for evaluating the performance of causal discovery methods. In this work, we evaluate causal discovery algorithms as binary classifiers deciding the presence or lack of n^2 edges among n nodes. This allows us to evaluate algorithms using standard classification metrics such as precision, recall, and F1 score (Davis and Goadrich, 2006; Fawcett, 2006; Powers, 2020; Sokolova and Lapalme, 2009; Tharwat, 2020). Given that the F1 score provides a balanced trade-off between precision and recall, we use it as our measure of accuracy. We define two separate metrics, (full) F1 score and adjacency F1 score. For the former, each of the n^2 edges (including any self-loops due to dampening autocorrelation for each node) in the graph is considered as one test sample for classification. In the latter, the ground-truth and learned graphs are first transformed into an undirected skeleton, placing an edge between two nodes if a directed edge existed in at least one direction. The resulting $\binom{n}{2}$ possible edges are then treated as test samples for classification and computation of adjacency F1 score.

To determine the degree and causal flow of a node i in a *binary* directed graph, its in-degree (number of edges pointing toward node i) and out-degree (number of edges originating from node i) are first computed and normalized by the total number of nodes in the graph. The degree of node i is then computed as the sum of the out-degree and in-degree, while the causal flow is obtained by subtracting the in-degree from the out-degree. The same process is followed for weighted graphs except that the calculation of in-degree and

out-degree involves a weighted mean. Mathematically,

$$Causal\ Flow\ (i) = \frac{1}{N}\sum_{j=1}^{N} G(i,j) - \frac{1}{N}\sum_{j=1}^{N} G(j,i) \quad , \quad i = 1, 2, ..., N$$

$$Degree\ (i) = \frac{1}{N}\sum_{j=1}^{N} G(i,j) + \frac{1}{N}\sum_{j=1}^{N} G(j,i) \quad , \quad i = 1, 2, ..., N$$

where G denotes the graph's (binary or weighted) adjacency matrix.

Computing Subnetwork Graphs from Parcel-Level Graphs

Subnetwork graphs were computed by aggregating parcel-level binary graphs into graphs between 16 subnetworks. These subnetworks consist of the standard 7 resting-state subnetworks (Yeo et al., 2011) plus one subcortical subnetwork, separately for the left and right hemispheres. A subnetwork-level graph is then computed for each subject, whereby the weight of an edge from subnetwork i to j is the number of nodes in subnetwork i that connect to nodes in subnetwork j, normalized by the number of all possible edges between these subnetworks. The results are then averaged over the subject, as depicted in Supplementary Figure 49.

Computing

All the computations reported in this study were performed on a Lenovo P620 workstation with AMD 3970X 32-Core processor, Nvidia GeForce RTX 2080 GPU, and 512GB of RAM.

Additional Information

Author Contributions

EN and AG designed and supervised the study; FA performed the research; HJ and MAKP assisted in the analyses of human fMRI data; FA and EN drafted and all authors edited the manuscript.

Acknowledgments

The research conducted in this study was partially supported by NSF Award #2239654 to EN, by the Canadian Institute for Advanced Research (fellowship awarded to MAKP), and by the Air Force Office of Scientific Research under award number FA9550-20-1-0106 (to MAKP).

Competing financial interests

The authors declare no competing financial interests.

Data Availability Statement

All the fMRI data used in this work is publicly available. The simulated fMRI benchmarks can be downloaded from `https://github.com/cabal-cmu/Feedback-Discovery` and the human fMRI data can be accessed via the HCP S1200 Release at `https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release`.

Code Availability Statement

The Python code for this study is publicly available at `https://github.com/nozarilab/2023Arab_CaLLTiF`.

References

Adams, A., Albin, S., Amunts, K., Asakawa, T., Bernard, A., Bjaalie, J. G., Chakli, K., Deshler, J. O., De Koninck, Y., Ebell, C. J., et al. (2020). International brain initiative: an innovative framework for coordinated global brain research efforts. *Neuron*, 105(2):212–216.

Agcaoglu, O., Miller, R., Mayer, A. R., Hugdahl, K., and Calhoun, V. D. (2015). Lateralization of resting state networks and relationship to age and gender. *Neuroimage*, 104:310–325.

Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A., and Lippert, T. (2016). The human brain project: creating a european research infrastructure to decode the human brain. *Neuron*, 92(3):574–581.

Andrews-Hanna, J. R., Smallwood, J., and Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the new York Academy of Sciences*, 1316(1):29–52.

Assaad, C. K., Devijver, E., and Gaussier, E. (2022). Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819.

Banks, S. J., Zhuang, X., Bayram, E., Bird, C., Cordes, D., Caldwell, J. Z., Cummings, J. L., Initiative, A. D. N., et al. (2018). Default mode network lateralization and memory in healthy aging and alzheimer's disease. *Journal of Alzheimer's Disease*, 66(3):1223–1234.

Barch, D. M. (2017). Resting-state functional connectivity in the human connectome project: Current status and relevance to understanding psychopathology. *Harvard Review of Psychiatry*, 25(5):209–217.

Barnett, L. and Seth, A. K. (2014). The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68.

Bressler, S. L. and Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences*, 14(6):277–290.

Bullmore, E. and Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349.

Burgess, G., Kandala, S., Nolan, D., Laumann, T., Power, J., Adeyemo, B., Harms, M., Petersen, S., and Barch, D. (2016). Evaluation of denoising strategies to address motion-correlated artifacts in resting-state functional magnetic resonance imaging data from the human connectome project. *Brain Connectivity*, 6(9):669–680.

Buxton, R. B., Wong, E. C., and Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864.

Chickering, D. M. (2002a). Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498.

Chickering, D. M. (2002b). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3):306–324.

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15.

Ding, M., Chen, Y., and Bressler, S. L. (2006). Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, pages 437–460.

Elam, J. (2020). Hcp Data Release Updates: Known Issues and Planned Fixes. `https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Release+Updates%3A+Known+Issues+and+Planned+fixes`. Last accessed May 13, 2020.

Entner, D. and Hoyer, P. O. (2010). On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128.

Essen, D. C. V., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., and Ugurbil, K. (2013). The wu-minn human connectome project: an overview. *NeuroImage*, 80:62–79.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47.

Fornito, A. and Bullmore, E. T. (2015). Connectomics: A new paradigm for understanding brain disease. *European Neuropsychopharmacology*, 25(5):733–748. Advances in neuroscience for improved mental health: an ECNP analysis of progress and prospects.

Friston, K., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302.

Friston, K., Moran, R., and Seth, A. K. (2013). Analysing connectivity with granger causality and dynamic causal modelling. *Current opinion in neurobiology*, 23(2):172–178.

Friston, K., Preller, K. H., Mathys, C., Cagnan, H., Heinzle, J., Razi, A., and Zeidman, P. (2019). Dynamic causal modelling revisited. *NeuroImage*, 199:730–744.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36.

Friston, K. J., Kahan, J., Biswal, B., and Razi, A. (2014). A dcm for resting state fmri. *NeuroImage*, 94:396–407.

Gerhardus, A. and Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc.

Ghassami, A., Yang, A., Kiyavash, N., and Zhang, K. (2020). Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, pages 3494–3504. PMLR.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Goebel, R., Roebroeck, A., Kim, D.-S., and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21(10):1251–1261.

Goense, J. B. and Logothetis, N. K. (2008). Neurophysiology of the bold fmri signal in awake monkeys. *Current Biology*, 18(9):631–640.

Granger, C. W. J. (1969). Investigatinag causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the national academy of sciences*, 100(1):253–258.

Hasan, U., Hossain, E., and Gani, M. O. (2023). A survey on causal discovery methods for temporal and non-temporal data. *arXiv preprint arXiv:2303.15027*.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243.

Henry, T. R. and Gates, K. M. (2017). Causal search procedures for fmri: review and suggestions. *Behaviormetrika*, 44:193–225.

Huettel, S. A., Song, A. W., and McCarthy, G. (2009). *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2nd edition.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731.

Jorgenson, L. A., Newsome, W. T., Anderson, D. J., Bargmann, C. I., Brown, E. N., Deisseroth, K., Donoghue, J. P., Hudson, K. L., Ling, G. S., MacLeish, P. R., et al. (2015). The brain initiative: developing technology to catalyse neuroscience discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140164.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*.

Liao, W., Mantini, D., Zhang, Z., Pan, Z., Ding, J., Gong, Q., Yang, Y., and Chen, H. (2010). Evaluating the effective connectivity of resting state networks using conditional granger causality. *Biological cybernetics*, 102:57–69.

Ljung, L. (1999). System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, pages 1–14.

Logothetis, N. K. (2008). What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878.

Malinsky, D. and Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In Le, T. D., Zhang, K., Kıcıman, E., Hyvärinen, A., and Liu, L., editors, *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, volume 92 of *Proceedings of Machine Learning Research*, pages 23–47. PMLR.

Markov, N. T., Ercsey-Ravasz, M. M., Ribeiro Gomes, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M.-A., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral cortex*, 24(1):17–36.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Conference on Uncertainty in Artificial Intelligence*.

Meek, C. (1997). *Graphical models: selecting causal and statistical models*. PhD thesis, Carnegie Mellon University.

Mengotti, P., Käsbauer, A.-S., Fink, G. R., and Vossel, S. (2020). Lateralization, functional specialization, and dysfunction of attentional networks. *Cortex*, 132:206–222.

Morton, B. E. (2020). Brain executive laterality and hemisity. *Personality Neuroscience*, 3:e10.

Nielsen, J. A., Zielinski, B. A., Ferguson, M. A., Lainhart, J. E., and Anderson, J. S. (2013). An evaluation of the left-brain vs. right-brain hypothesis with resting state functional connectivity magnetic resonance imaging. *PloS one*, 8(8):e71275.

Nozari, E., Bertolero, M. A., Stiso, J., Caciagli, L., Cornblath, E. J., He, X., Mahadevan, A. S., Pappas, G. J., and Bassett, D. S. (2023). Macroscopic resting-state brain dynamics are best described by linear models. *Nature Biomedical Engineering*.

Nozari, E., Pasqualetti, F., and Cortés, J. (2019). Heterogeneity of central nodes explains the benefits of time-varying control scheduling in complex dynamical networks. *Journal of Complex Networks*, 7(5):659–701.

Nunez, P. L. and Srinivasan, R. (2006). *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press.

Okano, H., Miyawaki, A., and Kasai, K. (2015). Brain/minds: brain-mapping project in japan. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140310.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Beaumont, P., Georgatzis, K., and Aragam, B. (2020). Dynotears: Structure learning from time-series data. *ArXiv*, abs/2002.00498.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Pearl, J. (2009a). Causal inference in statistics: An overview.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Poo, M.-m., Du, J.-l., Ip, N. Y., Xiong, Z.-Q., Xu, B., and Tan, T. (2016). China brain project: basic neuroscience, brain diseases, and brain-inspired computing. *Neuron*, 92(3):591–596.

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the national academy of sciences*, 98(2):676–682.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., and Glymour, C. (2010). Six problems for causal inference from fmri. *Neuro/mage*, 223:1545–1558.

Ramsey, J. D., Zhang, K., Glymour, M., Romero, R. S., Huang, B., Ebert-Uphoff, I., Samarasinghe, S., Barnes, E. A., and Glymour, C. (2018). Tetrad—a toolbox for causal discovery. In *8th international workshop on climate informatics*, pages 1–4.

Rawls, E., Kummerfeld, E., Mueller, B. A., Ma, S., and Zilverstand, A. (2022). The resting-state causal human connectome is characterized by hub connectivity of executive and attentional networks. *Neuroimage*, 255:119211.

Roebroeck, A., Formisano, E., and Goebel, R. (2005). Mapping directed influence over the brain using granger causality and fmri. *Neuroimage*, 25(1):230–242.

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996.

Sanchez-Romero, R., Ramsey, J., Zhang, K., Glymour, M., Huang, B., and Glymour, C. (2019). Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 86:274–306.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114.

Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297.

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41:65–98.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for fmri. *NeuroImage*, page 875–891.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 499–506.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer.

Stephan, K. E. and Roebroeck, A. (2012). A short history of causal modeling of fmri data. *Neuroimage*, 62(2):856–863.

Sun, X., Liu, G., Poupart, P., and Schulte, O. (2021). Nts-notears: Learning nonparametric temporal dags with time-series data and prior knowledge. *arXiv e-prints*, pages arXiv–2109.

Sutton, S. and Begleiter, H. (1979). Evoked brain potentials and behavior.

Sutton, S., Braren, M., Zubin, J., and John, E. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188.

Swanson, N., Eichele, T., Pearlson, G., Kiehl, K., Yu, Q., and Calhoun, V. D. (2011). Lateral differences in the default mode network in healthy controls and patients with schizophrenia. *Human brain mapping*, 32(4):654–664.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192.

Tian, Y., Margulies, D. S., Breakspear, M., and Zalesky, A. (2020). Hierarchical organization of the human subcortex unveiled with functional connectivity gradients. *bioRxiv*.

Tigges, J., Spatz, W., and Tigges, M. (1973). Reciprocal point-to-point connections between parastriate and striate cortex in the squirrel monkey (saimiri). *Journal of Comparative Neurology*, 148(4):481–489.

Vossel, S., Geng, J. J., and Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, 20(2):150–159.

Winder AT, Echagarruga C, Z. Q. and PJ, D. (2017). Weak correlations between hemodynamic signals and ongoing neural activity during the resting state. *Nature Neuroscience*, 18:17611769–640.

Yeo, T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Supplementary Material for "Whole-Brain Causal Discovery Using fMRI"

Fahimeh Arab[1], AmirEmad Ghassami[2], Hamidreza Jamalabadi[3], Megan A. K. Peters[4,5,6], and Erfan Nozari[1,7,8,*]

[1] *Department of Electrical and Computer Engineering, University of California, Riverside, USA*
[2] *Department of Mathematics and Statistics, Boston University, USA*
[3] *Department of Psychiatry and Psychotherapy, Phillips University of Marburg, Germany*
[4] *Department of Cognitive Sciences, University of California, Irvine, USA*
[5] *Center for the Neurobiology of Learning & Memory, University of California, Irvine, USA*
[6] *Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research, Canada*
[7] *Department of Mechanical Engineering, University of California, Riverside, USA*
[8] *Department of Bioengineering, University of California, Riverside, USA*
[*] *Corresponding author (email: erfan.nozari@ucr.edu)*

1 Supplementary Note 1: Contemporaneous Causal Effects

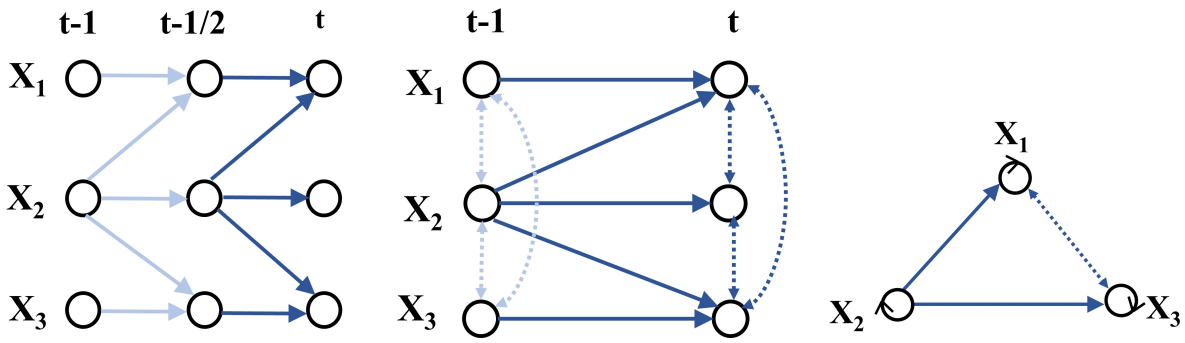
Contemporaneous causal connections play an important role in CaLLTiF and are one of its main advantages over several alternative algorithms such as MVGC and DCM. In this Supplementary Note, we define what we mean by a contemporaneous edge, explain why they arise as a result of temporal under-sampling, and provide further details on their processing and importance in CaLLTiF.

Unlike what the name may suggest, a contemporaneous edge does *not* mean or imply an instantaneous causal effect. In most physical systems, and certainly in the brain, causation takes time. However, ground-truth causation can occur at a timescale significantly faster than the sampling timescale of observations. Spike generation, axonal conduction, and synaptic transmission which constitute the backbone of causal flow in the brain, for example, take on the order of 10^1-10^2ms while fMRI sampling using state-of-the-art scanners takes on the order of 10^3ms. This mismatch between timescales effectively generates hidden variables, namely, the values of all of $X_1(t), \ldots, X_n(t)$ in between sampling times–e.g., $X_1(0.5)$, $X_2(1.2)$, etc.

This effect is illustrated in an example in Supplementary Figure 1. For the sake of simplicity, consider only 3 nodes and a ratio of 2-to-1 between the observation timescale and the ground-truth causation timescale. The variables $X_i(t-1/2), i = 1, 2, 3$ are all latent variables that can cause the observed variables $X_i(t), i = 1, 2, 3$ but, due to being latent, cannot be conditioned on. For instance, $X_1(t)$ and $X_2(t)$ are both caused by $X_2(t - 1/2)$, generating a statistically significant correlation between them that cannot be resolved by conditioning on $X_2(t - 1/2)$. Theoretically, this latent-confounded statistical dependence is what we define as a *contemporaneous edge* in CaLLTiF, which is the same as what is often shown by a bidirectional edge in an acyclic directed mixed graph (ADMG) (Richardson et al., 2023).

The meaning of a lagged edge in CaLLTiF should also be understood in the same context where slow sampling leads to latent nodes in between sampling times. Assuming that no direct (mono-synaptic) causation can take nearly as long as one TR (720ms in HCP and often longer in most fMRI recordings), a lagged edge in CaLLTiF corresponds to a directed *path* in the ground-truth causal graph in which all the intermediate nodes are latent. This is also the same as what is marked by a directed edge between measured variables in an ADMG. As an example, in Supplementary Figure 1, the lagged edge from $X_2(t-1)$ to $X_1(t)$ in the middle panel captures the directed path $X_2(t-1) \to X_2(t-1/2) \to X_1(t)$ in the left panel. Similarly, all other lagged edges in the middle panel correspond to directed, latent-mediate paths through $X_1(t-1/2)$, $X_2(t-1/2)$, or $X_3(t-1/2)$.

In its last step, CaLLTiF incorporates a heuristic step for summarization across lags, as shown in the right panel of Supplementary Figure 1. When no contemporaneous edge exists between a pair of variables X_i and X_j, this step is simply a disjunction (OR) across the lagged edges. However, when a contemporaneous edge exists, two scenarios may happen: (i) one or more lagged edges also exist, or (ii) no lagged edges exist between that pair of variables. In (i), we use the direction of the lagged edges to disambiguate the (intrinsically undirected) contemporaneous edge and then summarize across all lagged and contemporaneous edges using disjunction. While this can be seen as effectively ignoring the contemporaneous edge in building the summary graph, it can be very helpful in avoiding spurious two-cycles when only a uni-directional effect exists. This is the case, for example, between X_2 and X_1 (as well as X_2 and X_3) in Supplementary Figure 1. The lagged edge from X_2 to X_1 (and X_3) disambiguates the direction of the contemporaneous effect between them, leading to a correct unidirectional edge in the summary graph. In fact, it is straightforward to show that when all nodes have persistent auto-causation (the horizontal chains in the left panel of Supplementary Figure 1), the lack of any lagged edge from a node X_i to any other node X_j necessarily implies the lack of any direct causal effect from X_i to X_j. In other words, if lagged edges are only present in one direction, not putting an edge in the opposite direction in the summary graph is always a correct decision. Finally, if no such lagged edges are present to disambiguate the direction of the contemporaneous effect between a pair of variables (case (ii)), then a two-cycle is placed between the two variables in the summary graph which is the most likely prior for a pair of nodes in brain networks (Tigges et al., 1973; Felleman and Van Essen, 1991; Markov et al., 2014).



Supplementary Figure 1: Example showing how contemporaneous causal links are generated in CaLLTiF and how the final summary graph is constructed based on that. (a) Ground truth window causal graph in the true causal timescale (b) Corresponding window causal graph estimated by CaLLTiF (c) Corresponding summary causal graph outputted by CaLLTiF

A note is here warranted on why we need contemporaneous edges at all. Theoretically, if both a contemporaneous edge and one or more lagged edges exist between a pair of variables (case (i)), then the contemporaneous edge is (and should be) effectively ignored, as described above. On the other hand, if only a contemporaneous edge exists between two nodes X_i and X_j (case (ii)), it can be shown under the same assumption of persistent auto-causation above that the contemporaneous edge can only be due to a latent common cause $X_k(t-\tau)$ where $k \neq i, j$ and $t-1 < \tau < t$. In other words, the contemporaneous edge is spurious in this case. An example of this can be seen between X_1 and X_3 in Supplementary Figure 1. As such, in both cases, the contemporaneous effect seems either ineffective or wrong. This is true, however, only in theory when one has sufficient data, computational power, and knowledge of the exact parametric forms of the causal relationships between all variables to detect any and all lagged effects, no matter how weak or noisy they may be.

In reality, however, noise is constantly injected at every time point and statistical correlations between variables at time $t-1$ and those at time t decay exponentially with the number of unobserved intermediate steps between them. As a result, with the typical number of data samples and timescale differences in fMRI, causal relationships between a vast majority of pairs of variables can only be detected contemporaneously. This can be clearly seen from Figure 6a in the main text, where about %70 of detectable effects are contemporaneous only. This is the case even with the "fast" sampling time of 0.72s in HCP data. In most other state-of-the-art fMRI recordings, sampling times are between 1-3s, making contemporaneous effects almost all that remain detectable.

2 Supplementary Note 2: Distinctions Between CaLLTiF and MVGC

Similar to MVGC, CaLLTiF detects the presence of an edge using multivariate linear autoregressive models and statistical hypothesis testing to determine statistical significance. In this Supplementary Note, we discuss their differences in further detail and explain why CaLLTiF can achieve better accuracy in large networks, such as the simulated Full Macaque and HCP datasets we examined in this work. For ease of reference and uniformity of notation, we first revisit the underlying statistical tests used by each algorithm.

2.1 Statistical test for determining the presence of an edge in CaLLTiF

Computing the test statistic (partial correlation coefficient). To determine the presence of an edge between $X_i(t-\tau)$ and $X_j(t)$ in CaLLTiF, we first regress out the effects of all potential common causes and mediators from both variables. This is done by constructing two linear regression models, one that predicts $X_j(t)$ and another that predicts $X_i(t-\tau)$ from all potential common causes and mediators.

$$X_i(t-\tau) = \sum_{\substack{k\in\{1,\ldots,N\}\\ s\in\{1,\ldots,\tau_{\max}\}\\ (k,s)\neq(i,\tau)}} \alpha_{k,s}X_k(t-s) + E_{X_i}(t), \qquad t=\tau_{\max},\ldots,T-1 \tag{S1a}$$

$$X_j(t) = \sum_{\substack{k\in\{1,\ldots,N\}\\ s\in\{1,\ldots,\tau_{\max}\}\\ (k,s)\neq(i,\tau)}} \beta_{k,s}X_k(t-s) + E_{X_j}(t), \qquad t=\tau_{\max},\ldots,T-1 \tag{S1b}$$

The regression coefficients $\{\alpha_{k,s}\}$ and $\{\beta_{k,s}\}$ are fit using standard least squares, the residuals of each model are calculated,

$$\hat{E}_{X_i}(t) = X_i(t-\tau) - \sum_{\substack{k\in\{1,\ldots,N\}\\ s\in\{1,\ldots,\tau_{\max}\}\\ (k,s)\neq(i,\tau)}} \hat{\alpha}_{k,s}X_k(t-s), \qquad t=\tau_{\max},\ldots,T-1 \tag{S2a}$$

$$\hat{E}_{X_j}(t) = X_j(t) - \sum_{\substack{k\in\{1,\ldots,N\}\\ s\in\{1,\ldots,\tau_{\max}\}\\ (k,s)\neq(i,\tau)}} \hat{\beta}_{k,s}X_k(t-s), \qquad t=\tau_{\max},\ldots,T-1 \tag{S2b}$$

and the Pearson correlation coefficient is calculated between $\hat{E}_{X_i}(t)$ and $\hat{E}_{X_j}(t)$. Let $r_{i,j,\tau}$ denote this correlation coefficient, a.k.a. the partial correlation coefficient between $X_i(t-\tau)$ and $X_j(t)$ conditioned on $\{X_k(t-s) : (k,s)\neq(i,\tau)\}$.

Hypothesis testing (t-test). We place an edge between $X_i(t-\tau)$ and $X_j(t)$ if the partial correlation coefficient $r_{i,j,\tau}$ is statistically significantly different from 0. This is done using the standard t-test for Pearson correlation coefficient, namely, with the null hypothesis $H_0 : r_{i,j,\tau} = 0$ and the t statistic

$$t = \begin{cases} \frac{r_{i,j,\tau}}{\sqrt{1-r_{i,j,\tau}^2}}\sqrt{T-N\tau_{\max}-2} & \text{if} \quad \tau = 0 \\ \frac{r_{i,j,\tau}}{\sqrt{1-r_{i,j,\tau}^2}}\sqrt{T-N\tau_{\max}-1} & \text{if} \quad \tau > 0 \end{cases}$$

where T is the number of observations, N is the number of variables, and $\tau_{\max}$ is the number of time lags.

2.2 Statistical test for determining the presence of an edge in MVGC

Computing the test statistic (conditional Granger causality index, CGCI). To determine the presence of an edge from X_i to X_j, MVGC also fits two linear regression models (note the lack of time indexing, as MVGC learns the summary graph directly instead of learning an extended window graph and summarizing it as done in CaLLTiF). A "full model" predicts $X_j(t)$ using the past history of all the variables, while a "restricted model" predicts $X_j(t)$ using the past history of all other variables but X_i:

Full model: $$X_j(t) = \sum_{\substack{k\in\{1,\ldots,N\}\\ s\in\{1,\ldots,\tau_{\max}\}}} \alpha_{k,s}X_k(t-s) + E_{X_j}(t), \qquad t=\tau_{\max},\ldots,T-1 \tag{S3a}$$

Restricted model: $$X_j(t) = \sum_{\substack{k\in\{1,\ldots,N\}\setminus\{i\}\\ s\in\{1,\ldots,\tau_{\max}\}}} \beta_{k,s}X_k(t-s) + E_{X_j\setminus X_i}(t), \qquad t=\tau_{\max},\ldots,T-1 \tag{S3b}$$

Estimated residuals $\hat{E}_{X_j}$ and $\hat{E}_{X_j \setminus X_i}$ are then computed similar to Eq. (S2) and the CGCI from X_i to X_j is defined as:

$$CGCI_{X_i \to X_j} = \ln \left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_F^2} \right)$$

where $\hat{\sigma}_F^2$ and $\hat{\sigma}_R^2$ are the variances of $\hat{E}_{X_j}$ and $\hat{E}_{X_j \setminus X_i}$, respectively.

Hypothesis testing (F-test). The statistical significance of the CGCI is tested using an F-test with the statistic

$$F = \frac{(\hat{\sigma}_R^2 - \hat{\sigma}_F^2)/\tau_{\max}}{\hat{\sigma}_F^2/(T - \tau_{\max} - N\tau_{\max})} = \frac{T - \tau_{\max} - N\tau_{\max}}{\tau_{\max}} \left(e^{CGCI_{X_i \to X_j}} - 1 \right)$$

where T is the number of observations and $N\tau_{\max}$ is the number of coefficients in the full model.

2.3 Comparison between independence test in MVGC and CaLLTiF

The strongest similarity between MVGC and CaLLTiF comes from their underlying linear regression models in Eq. (S3) and Eq. (S1). Nevertheless, there are major differences in how these regression models are used and how their statistical significance is assessed in each algorithm. In particular, partial correlation measures the *remaining association* between two variables *after having removed* the effect of all others. On the other hand, CGCI measures the *added predictive power* of the history of one variable in explaining the other, while all other variables are kept as regressors in a multivariate system. In other words, both algorithms are based on how much one variable (X_i) can explain variance in another (X_j), but CaLLTiF assesses the significance of this relationship *per se* (in an *absolute* sense), while MVGC compares the additional explained variance coming from X_i with how much variance all other variables were unable to explain (in a *relative* sense).

This difference between how regression models are used in MVGC and CaLLTiF can lead to a higher sensitivity to the total number of variables in MVGC. In our experiments both with simulated and real data ($N \sim 100$), we indeed found MVGC to generate overly sparse graphs for standard significance thresholds ($\alpha \leq 0.05$) and unreliable/overly dense graphs for higher thresholds. Specifically, for the Full Macaque simulated dataset, we swept over the significance threshold of MVGC (like we did for all other algorithms) and found its highest F1 score (as shown in Figure 3 in the main text). This was achieved at $\alpha = 0.5$, i.e., when the algorithm returns a complete graph. In other words, any value of α lower than 0.5 led to such inaccurate removal of edges that the result was worse than the complete graph. Likewise, in our experiments with the HCP data, even though we can no longer compute or optimize the ground-truth accuracy of MVGC over its significance threshold, we ran MVGC at the same significance level of 0.01 used for CaLLTiF. The resulting graphs, however, were completely empty. This is even despite the fact that CaLLTiF actually uses a much smaller significance threshold (0.01/32) for each t-test due to its conservative correction for multiple comparisons across lags. In other words, while CaLLTiF can find hundreds of edges with %99 or higher statistical confidence, MVGC can find none. One can certainly increase the significance threshold and have MVGC detect more edges, but at the risk of increased false positives (type I error) when there is no ground-truth to compare the results against.

3 Supplementary Note 3: On the Rationale for Only Conditioning on Past Values in CaLLTiF

A key element in CaLLTiF (and constraint-based methods for causal discovery in general) that allows for distinguishing between correlation and causation is conditioning on variables that can potentially be a common cause or mediator. The fork ($X_1 \leftarrow X_2 \rightarrow X_3$) and chain ($X_1 \rightarrow X_2 \rightarrow X_3$) architectures are the simplest network motifs that showcase the effect of a common cause and mediator, respectively. In the fork motif, X_1 and X_3 are correlated but neither causes the other–their correlation is due to their common cause X_2. Assuming linear Gaussian generative mechanisms, it is straightforward to show that the Pearson correlation coefficient between X_1 and X_3 is statistically significantly nonzero, while their partial correlation coefficient after conditioning on X_2 is zero on average (finite samples) and asymptotically (infinite samples). Similarly in the chain motif, conditioning on X_2 makes X_1 and X_3 statistically independent, as desired.

Nevertheless, conditioning is not universally beneficial. Excessive conditioning can also generate spurious correlations (and therefore edges) between variables that are otherwise not correlated. The simplest network motif showcasing this effect is the collider, a.k.a. V-structure ($X_1 \rightarrow X_2 \leftarrow X_3$). In this network, X_1 and X_3 are independent per se, but they become dependent if one conditions on their common child X_2. Intuitively, knowing the value of X_2 can make certain combinations of X_1 and X_3 more or less likely to have occurred than their default, thus coupling their distributions. This is the reason why constraint-based methods for causal discovery, such as the celebrated PC algorithm (Spirtes and Glymour, 1991) and its variants (including PCMCI), spend significant computational resources on selecting the best conditioning set for each pair of variables.

In this work, we showed that conditioning on all past values of all recorded variables is not only computationally vital for scalability but also most accurate in simulated fMRI (Figure 16 in the main text). Nevertheless, given the importance of contemporaneous effects discussed here, one may wonder why conditioning is restricted to lagged variables ($s > 0$ in Eq. (S1)). In the following, we provide empirical evidence for why conditioning on contemporaneous variables can generate more spurious edges than it can prune, particularly in certain parametric regimes.

We generated linear Gaussian time series from the structural equation models (SEMs) of two network architectures: the chain ($X_1(t) \rightarrow X_2(t) \rightarrow X_3(t)$) and the collider ($X_1(t) \rightarrow X_2(t) \leftarrow X_3(t)$). Auto-causation was added for all variables to generate temporal auto-correlations–the strongest source of dynamics in neural data. For the chain network, the SEM takes the form

$$X_1(t) = \alpha X_1(t-1) + e_1(t) \tag{S4a}$$
$$X_2(t) = \alpha X_2(t-1) + \beta X_1(t) + e_2(t) \tag{S4b}$$
$$X_3(t) = \alpha X_3(t-1) + \beta X_2(t) + e_3(t) \tag{S4c}$$

while for the collider, the SEM reads

$$X_1(t) = \alpha X_1(t-1) + e_1(t) \tag{S5a}$$
$$X_2(t) = \alpha X_2(t-1) + \beta X_1(t) + \beta X_3(t) + e_2(t) \tag{S5b}$$
$$X_3(t) = \alpha X_3(t-1) + e_3(t) \tag{S5c}$$

All auto- and cross-causation weights are set uniformly at α and β, respectively, to simplify parametric analyses.

We then investigated the impact of conditioning solely on past values of X_2 compared to conditioning on both its past and present values on the partial correlation between $X_1(t)$ and $X_3(t)$. For the chain, we would ideally like conditioning to *maximally reduce* correlation between $X_1(t)$ and $X_3(t)$ compared to their unconditional (baseline) correlation coefficient. In contrast, for the collider, we ideally want conditioning *not to increase* the correlation between $X_1(t)$ and $X_3(t)$ compared to their unconditional baseline. Therefore,

we compare three correlation coefficients for each network motif: unconditional correlation ($R_{X_1(t),X_3(t)}$), partial correlation conditioning only on the past ($R_{X_1(t),X_3(t)|X_2(t-1),\ldots,X_2(t-\tau_{max})}$), and partial correlation conditioning on past and present ($R_{X_1(t),X_3(t)|X_2(t),X_2(t-1),\ldots,X_2(t-\tau_{max})}$).

Not surprisingly, $R_{X_1(t),X_3(t)|X_2(t-1),\ldots,X_2(t-\tau_{max})}$ almost always falls in between the other two, i.e., both the desired and undesired effects of conditioning only on the past are smaller compared to conditioning on past and present. Therefore, to assess whether conditioning only on the past (as done on CaLLTiF) is better than conditioning on the past and present overall, we computed the following metric for each network motif:

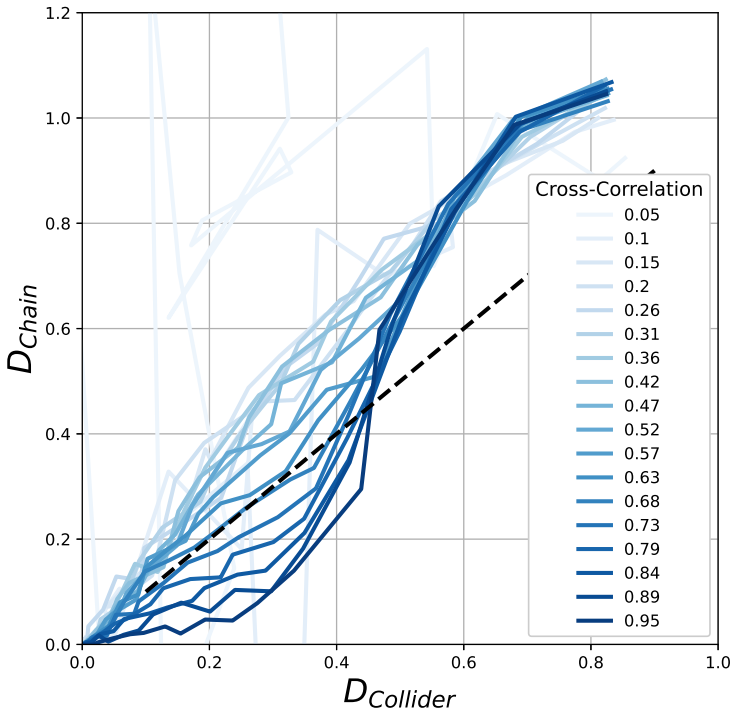
$$D = \frac{R_{X_1(t),X_3(t)|X_2(t-1),\ldots,X_2(t-\tau_{max})} - R_{X_1(t),X_3(t)}}{R_{X_1(t),X_3(t)|X_2(t),X_2(t-1),\ldots,X_2(t-\tau_{max})} - R_{X_1(t),X_3(t)}} \tag{S6}$$

This is simply a linear mapping of the interval between $R_{X_1(t),X_3(t)}$ and $R_{X_1(t),X_3(t)|X_2(t),X_2(t-1),\ldots,X_2(t-\tau_{max})}$ to $[0, 1]$, such that

- if $R_{X_1(t),X_3(t)|X_2(t-1),\ldots,X_2(t-\tau_{max})} = R_{X_1(t),X_3(t)}$ then $D = 0$, and
- if $R_{X_1(t),X_3(t)|X_2(t-1),\ldots,X_2(t-\tau_{max})} = R_{X_1(t),X_3(t)|X_2(t),X_2(t-1),\ldots,X_2(t-\tau_{max})}$ then $D = 1$.

Ideally, we would like $D = 1$ for the chain and $D = 0$ for the collider. In practice, conditioning on the past is better whenever $D_{\text{chain}} > D_{\text{collider}}$ while conditioning on past and present is preferred when $D_{\text{chain}} < D_{\text{collider}}$.

Supplementary Figure 2 shows the values of D_{chain} and D_{collider} for different values of SEM parameters α and β. Each line corresponds to one value of β (cross-correlation) and it is traversed from left to right as α is increased from 0 to 1. The area above the dashed black line indicates the range of parameters where conditioning only on the past is preferable ($D_{\text{chain}} > D_{\text{collider}}$). This is the case for all values of α if $\beta \lesssim 0.7$ (any autocorrelation strength in weakly- to moderate-connected networks) and for larger values of α if $\beta \gtrsim 0.7$ (strong autocorrelations in strongly-connected networks). This includes the majority of the joint values of (α, β), including brain networks for which regional autocorrelations are often notably stronger than correlations between regions (Nozari et al., 2023).



Supplementary Figure 2: The desired and undesired effects of conditioning only on past variables, as opposed to conditioning on both past and contemporaneous variables in simple network motifs. Each line shows the values of the metric D defined in Eq. (S6) for two simple network motifs: the chain ($X_1(t) \to X_2(t) \to X_3(t)$) for which conditioning is generally helpful by avoiding/weakening a spurious edge between X_1 and X_3, and the collider ($X_1(t) \to X_2(t) \leftarrow X_3(t)$) for which conditioning is generally harmful by creating a spurious edge between the two. Simulated time series data is generated from each network using the SEMs in Eq. (S4) and Eq. (S5). Each line corresponds to one value of β (cross-correlation) and it is traversed from left to right as α is increased from 0 to 1. The area above the dashed black line indicates the range of parameters where conditioning only on the past is preferable ($D_{\text{chain}} > D_{\text{collider}}$).

Supplementary References

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47.

Markov, N. T., Ercsey-Ravasz, M. M., Ribeiro Gomes, A., Lamy, C., Magrou, L., Vezoli, J., Misery, P., Falchier, A., Quilodran, R., Gariel, M.-A., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral cortex*, 24(1):17–36.

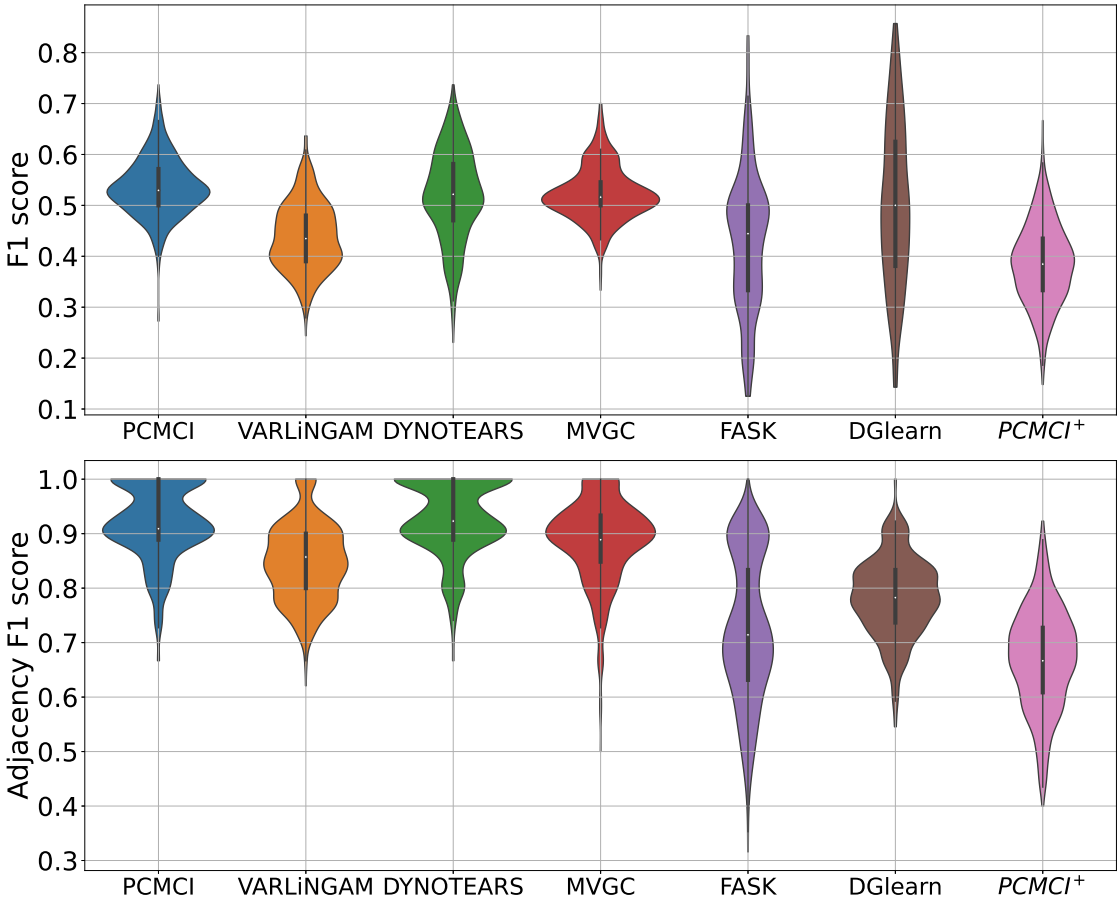
Nozari, E., Bertolero, M. A., Stiso, J., Caciagli, L., Cornblath, E. J., He, X., Mahadevan, A. S., Pappas, G. J., and Bassett, D. S. (2023). Macroscopic resting-state brain dynamics are best described by linear models. *Nature Biomedical Engineering*.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2023). Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361.

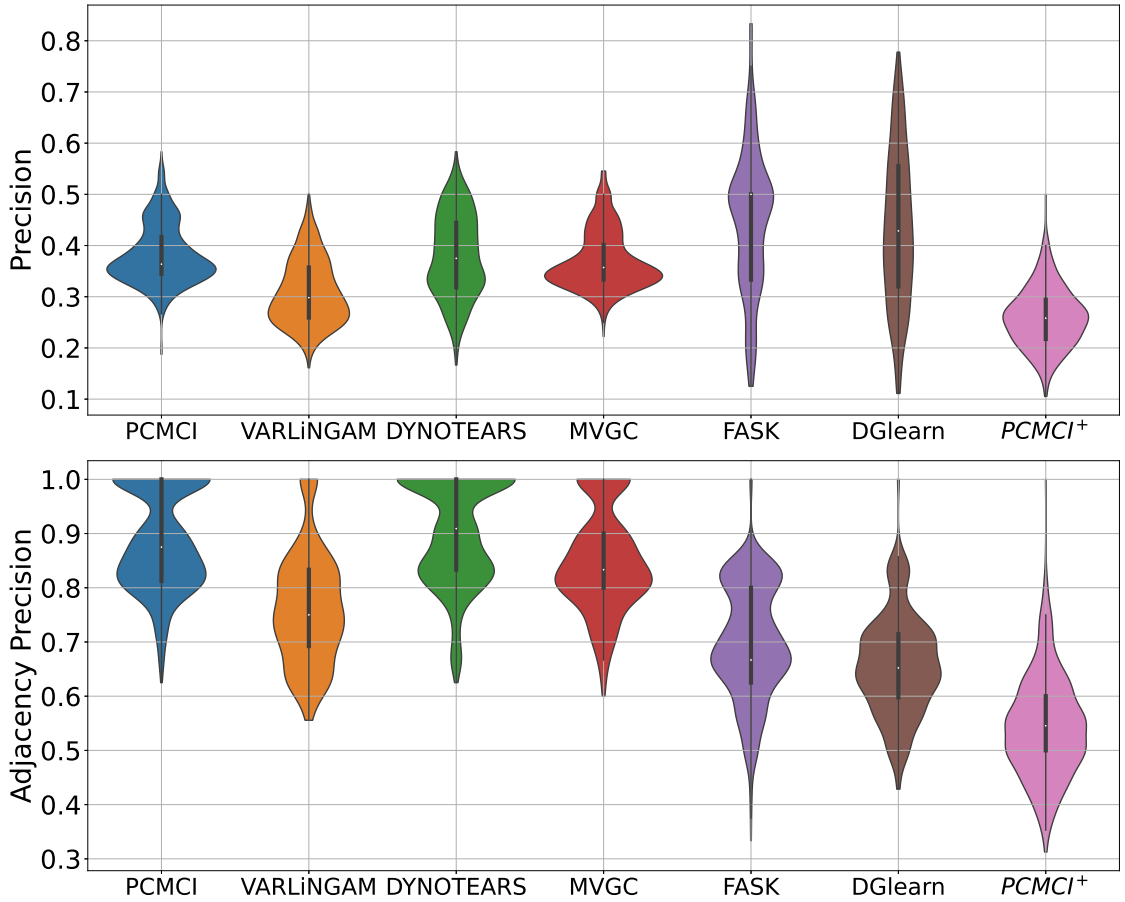
Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Tigges, J., Spatz, W., and Tigges, M. (1973). Reciprocal point-to-point connections between parastriate and striate cortex in the squirrel monkey (saimiri). *Journal of Comparative Neurology*, 148(4):481–489.

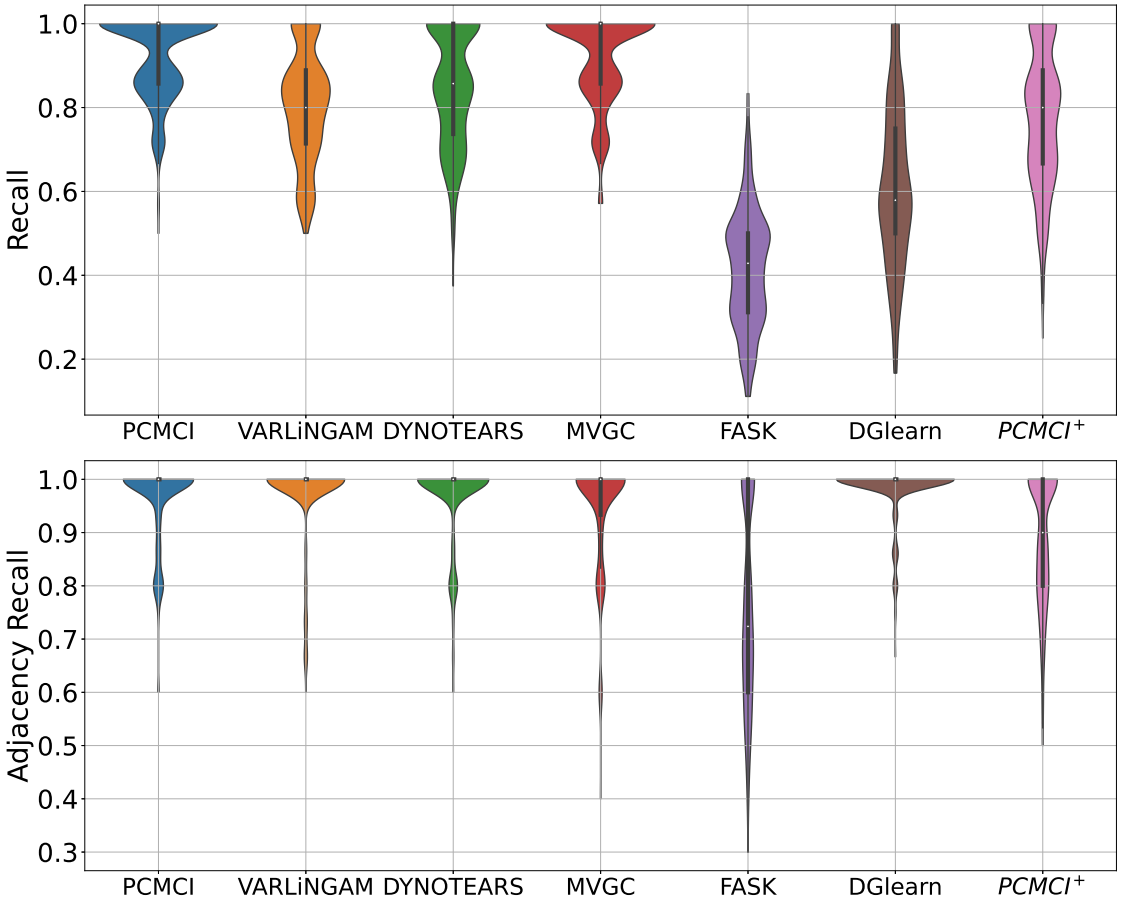
Supplementary Figures for Simulated fMRI from Simple Networks



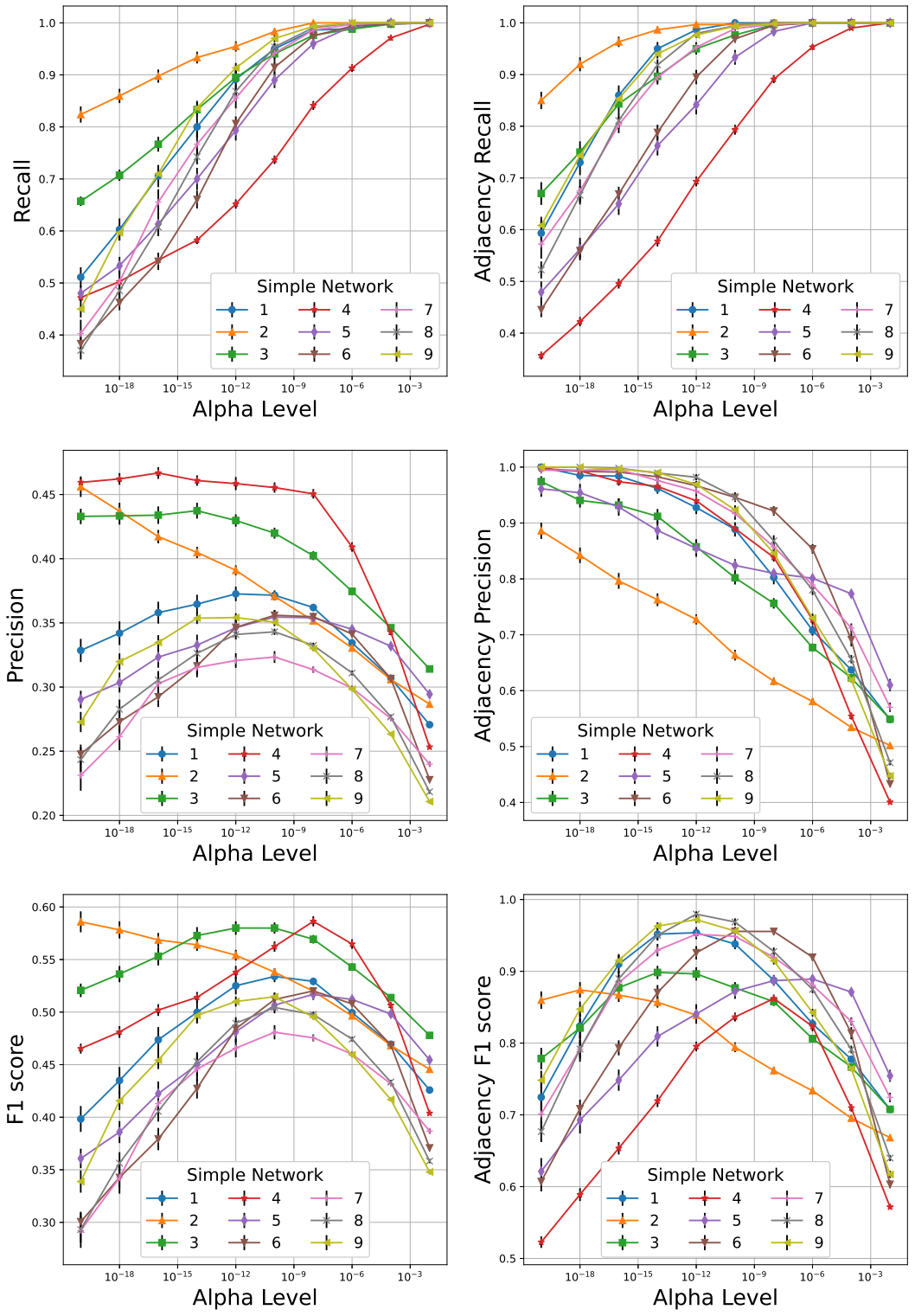
Supplementary Figure 3: Results of comparing several state-of-the-art causal discovery algorithms over a benchmark of simulated fMRI generated from simple networks with 5-10 nodes each. F1 score of seven state-of-the-art algorithms (six from Table 1 and MVGC) for correctly identifying the full (directed) graphs and corresponding undirected skeletons, respectively. All methods are evaluated using optimized values of their respective hyperparameters (see Methods). The benchmark data includes 60 repetitions of fMRI data from each of the 9 graphs, so each violin plot is based on 540 samples.



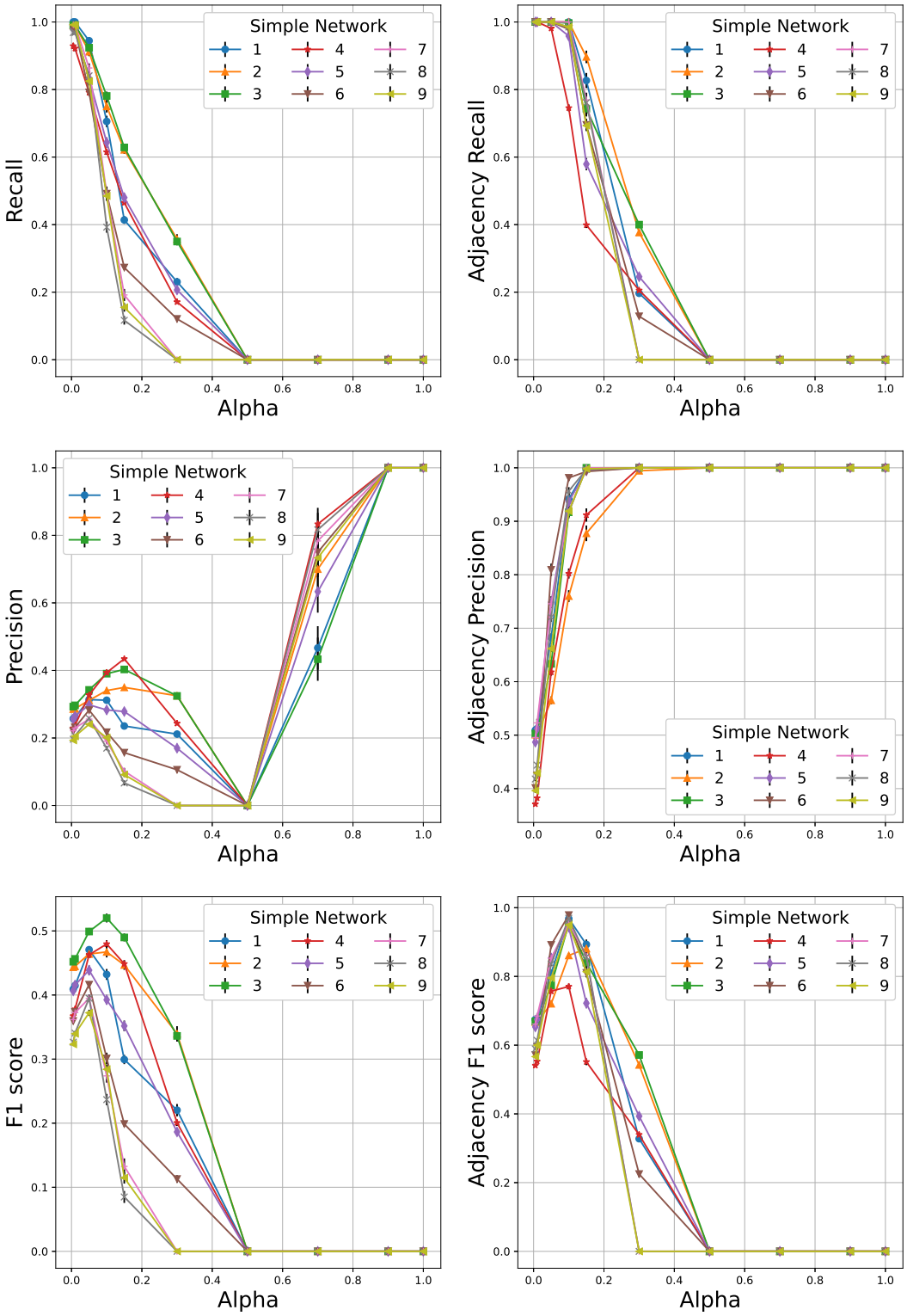
Supplementary Figure 4: Results of comparing several state-of-the-art causal discovery algorithms over a benchmark of simulated fMRI generated from simple networks with 5-10 nodes each. Precision of seven state-of-the-art algorithms (six from Table 1 and MVGC) for correctly identifying the full (directed) graphs and corresponding undirected skeletons, respectively. All methods are evaluated using optimized values of their respective hyperparameters (see Methods). The benchmark data includes 60 repetitions of fMRI data from each of the 9 graphs, so each violin plot is based on 540 samples.



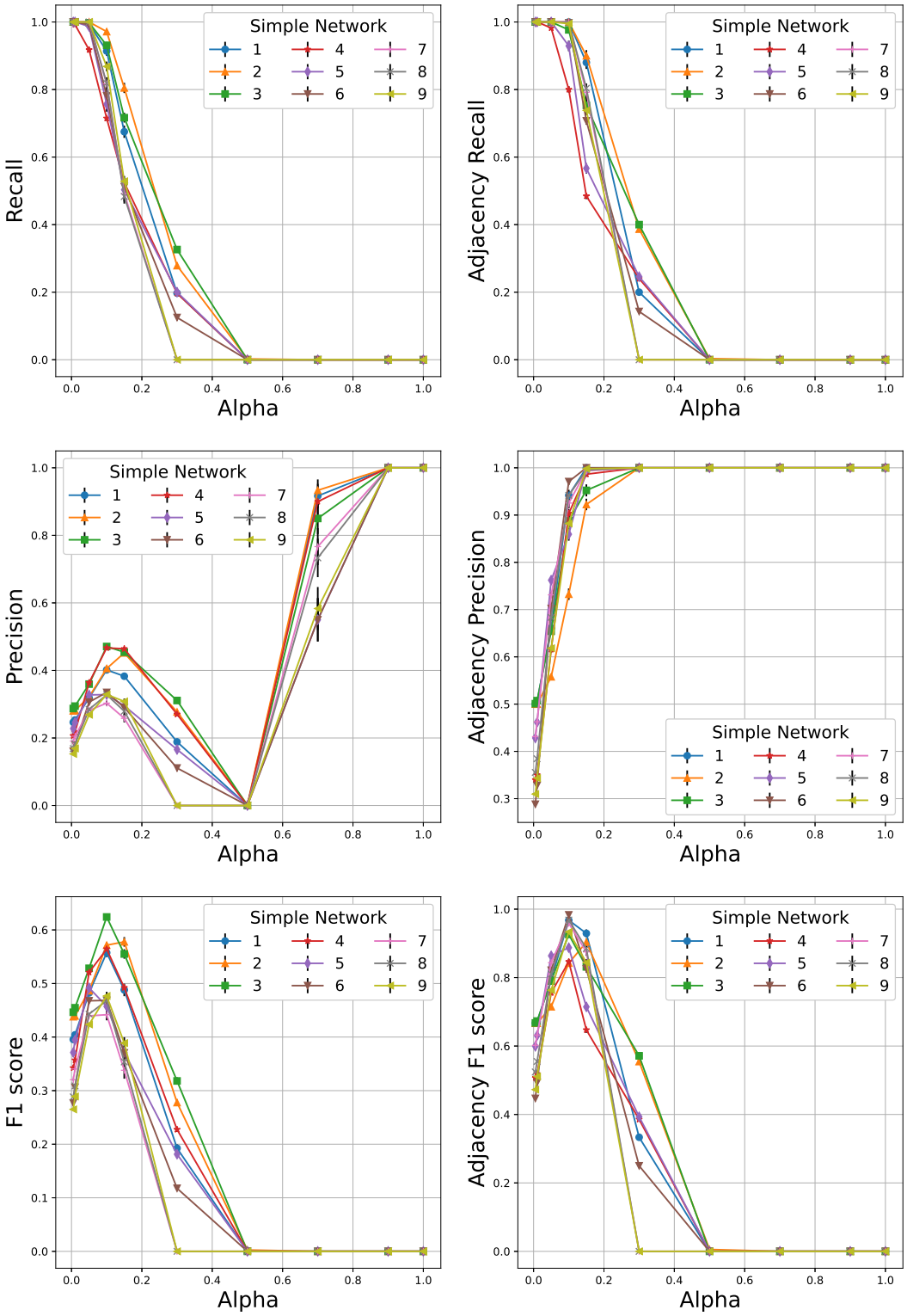
Supplementary Figure 5: Results of comparing several state-of-the-art causal discovery algorithms over a benchmark of simulated fMRI generated from simple networks with 5-10 nodes each. Recall of seven state-of-the-art algorithms (six from Table 1 and MVGC) for correctly identifying the full (directed) graphs and corresponding undirected skeletons, respectively. All methods are evaluated using optimized values of their respective hyperparameters (see Methods). The benchmark data includes 60 repetitions of fMRI data from each of the 9 graphs, so each violin plot is based on 540 samples.



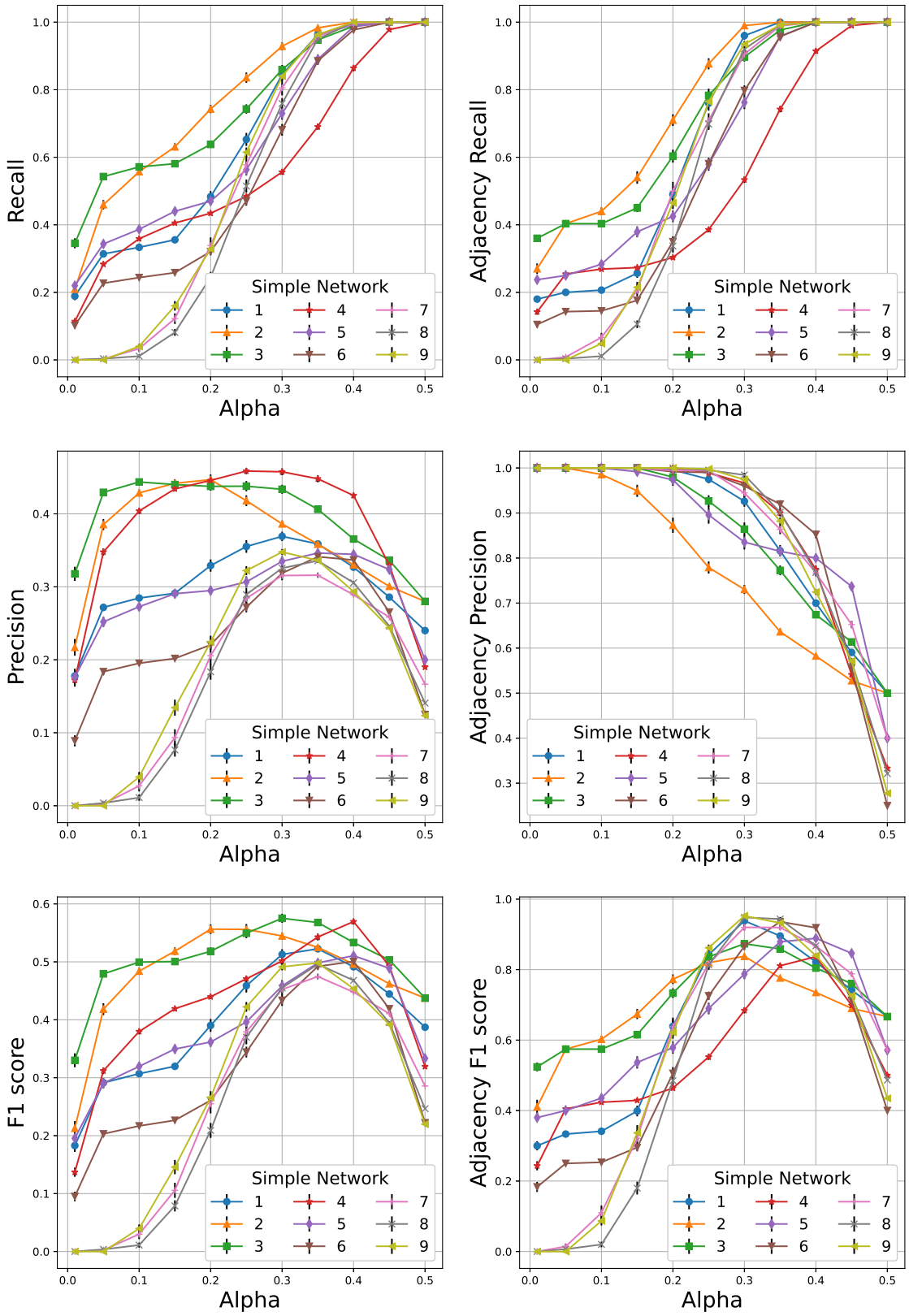
Supplementary Figure 6: Detailed performance curves of PCMCI over simulated fMRI from simple networks for varying values of its hyperparameter Alpha Level. In all graphs, the error bars depict the standard error of the mean.



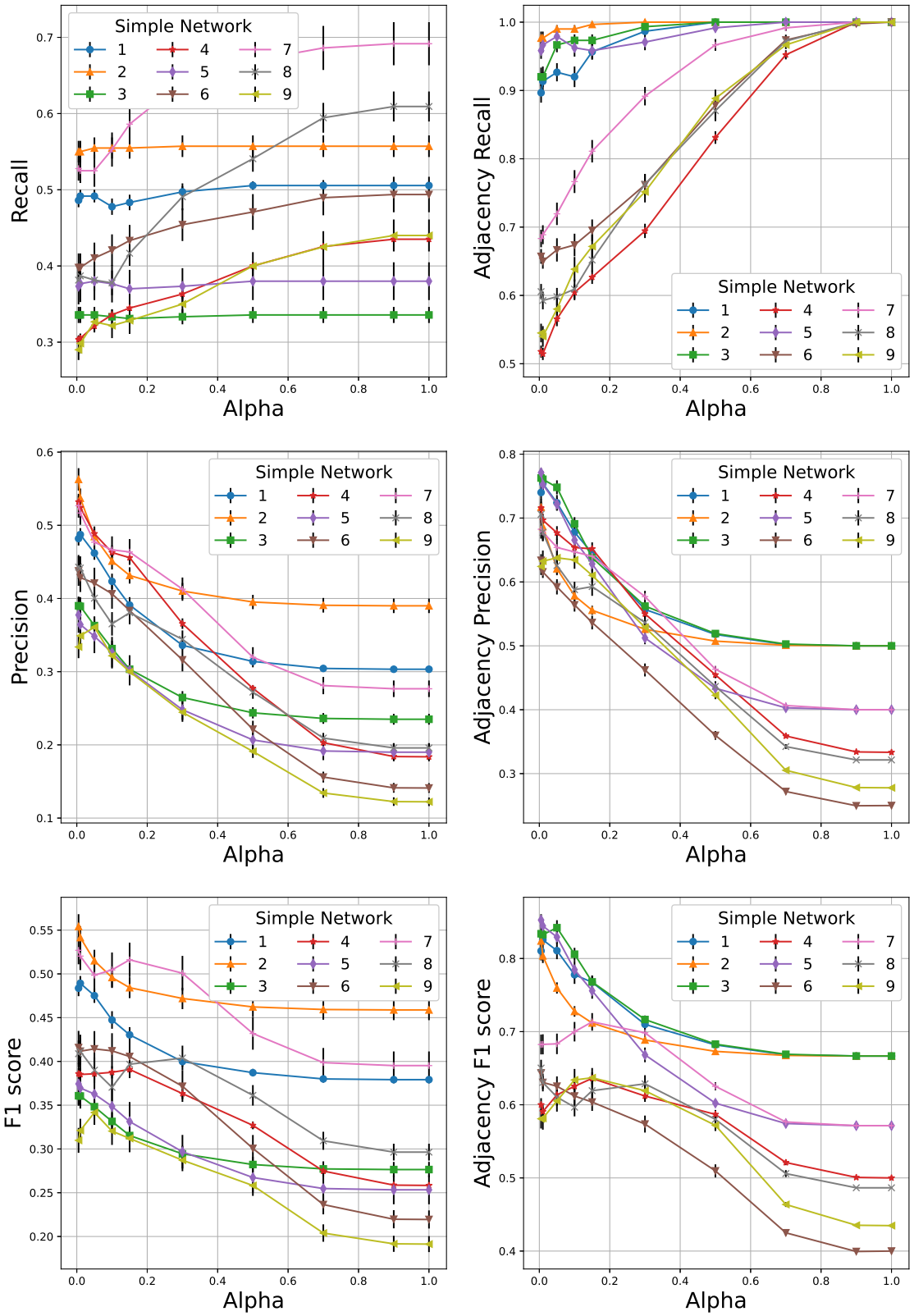
Supplementary Figure 7: Detailed performance curves of VARLiNGAM over simulated fMRI from simple networks for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



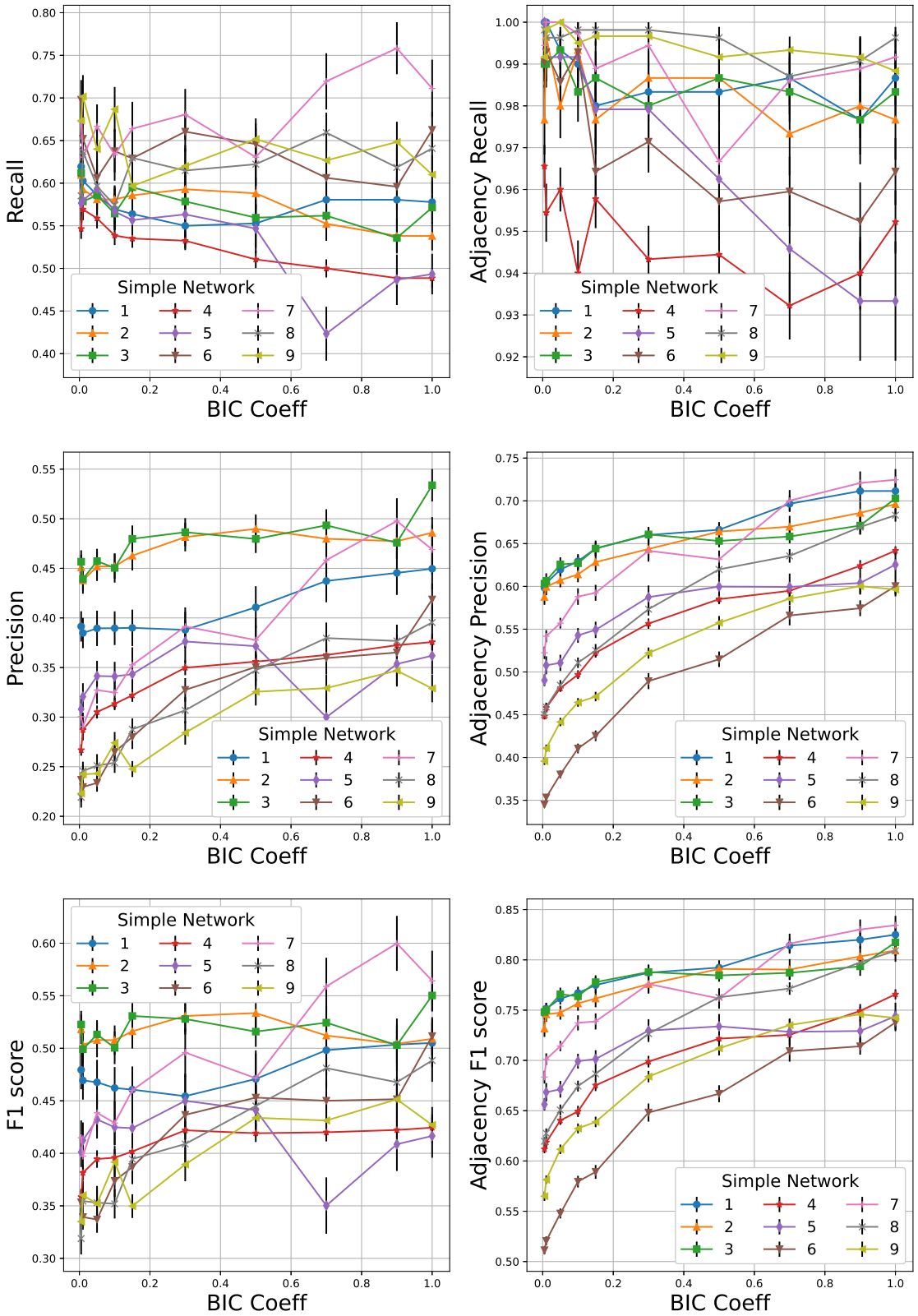
Supplementary Figure 8: Detailed performance curves of DYNOTEARS over simulated fMRI from simple networks for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



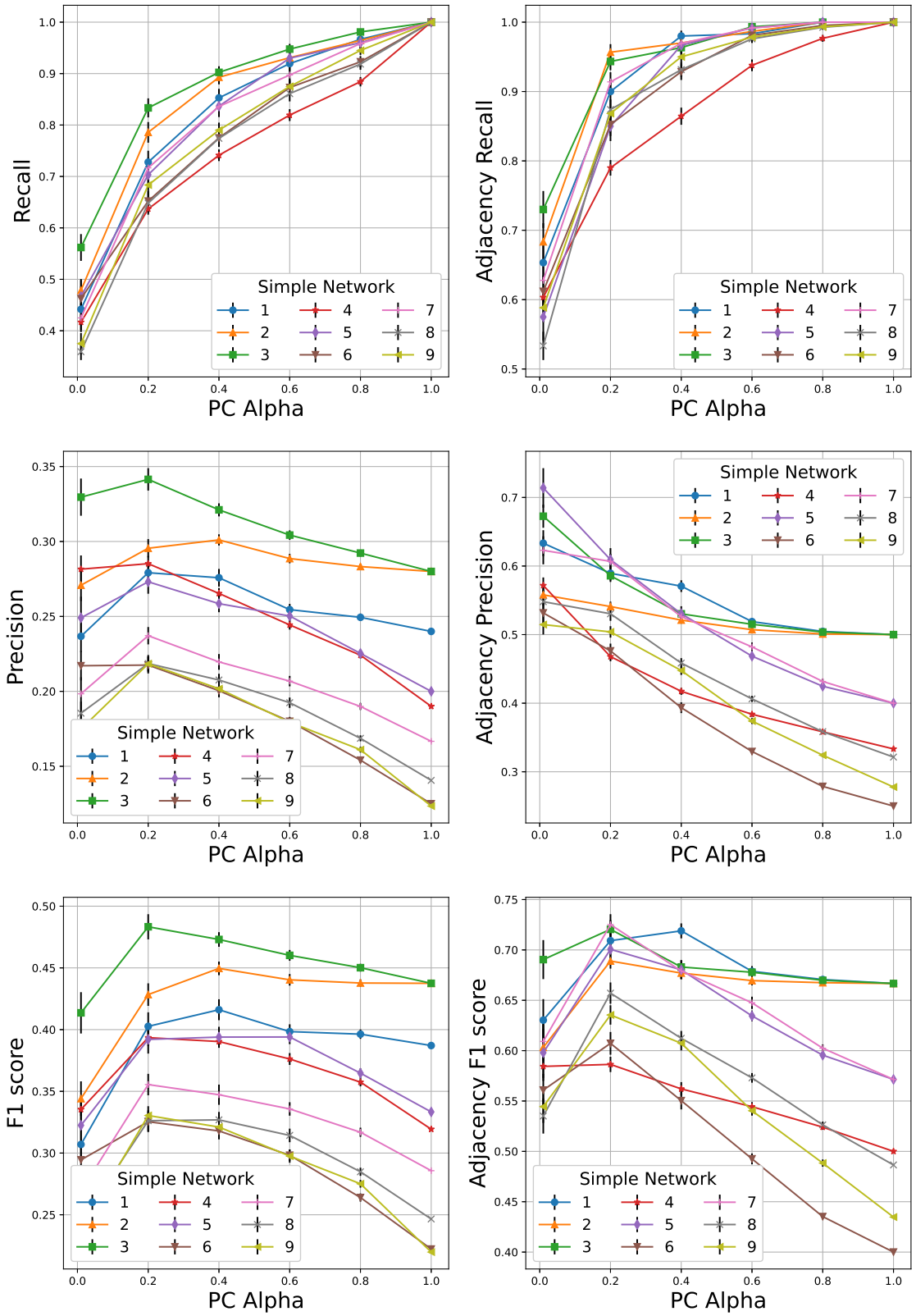
Supplementary Figure 9: Detailed performance curves of MVGC over simulated fMRI from simple networks for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



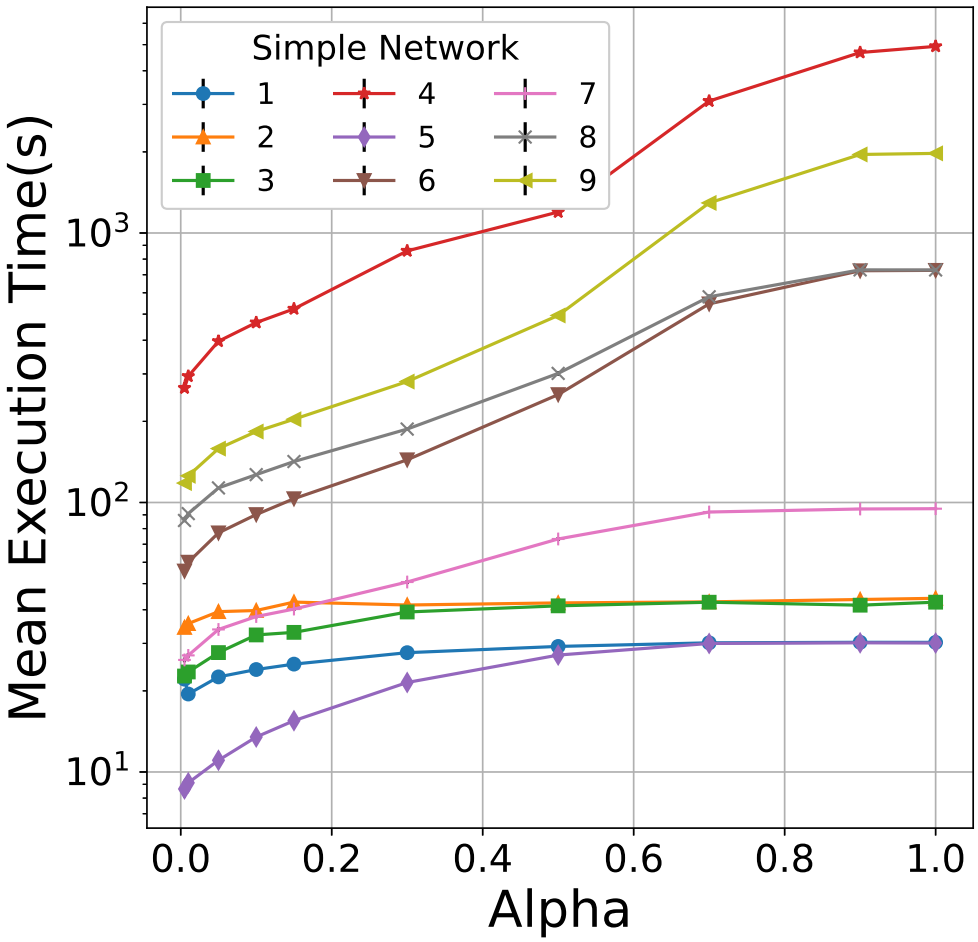
Supplementary Figure 10: Detailed performance curves of FASK over simulated fMRI from simple networks for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



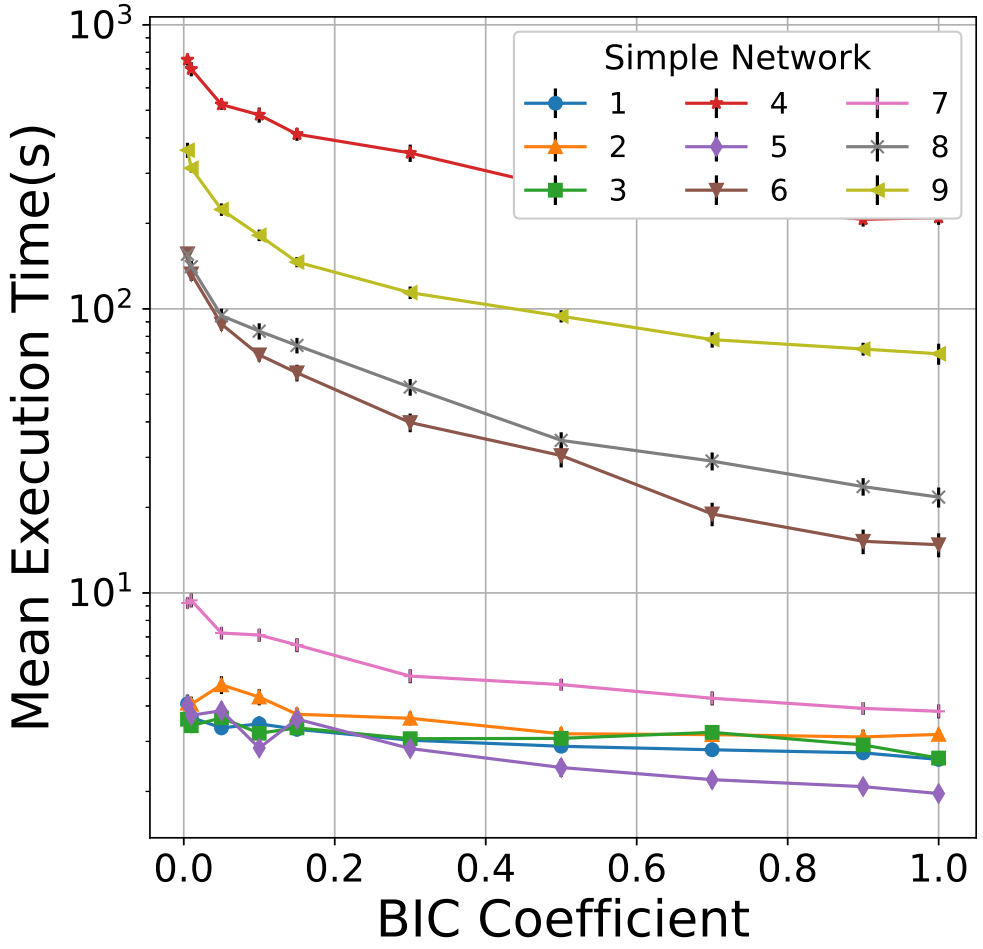
Supplementary Figure 11: Detailed performance curves of DGlearn over simulated fMRI from simple networks for varying values of its hyperparameter BIC coefficient. In all graphs, the error bars depict the standard error of the mean.



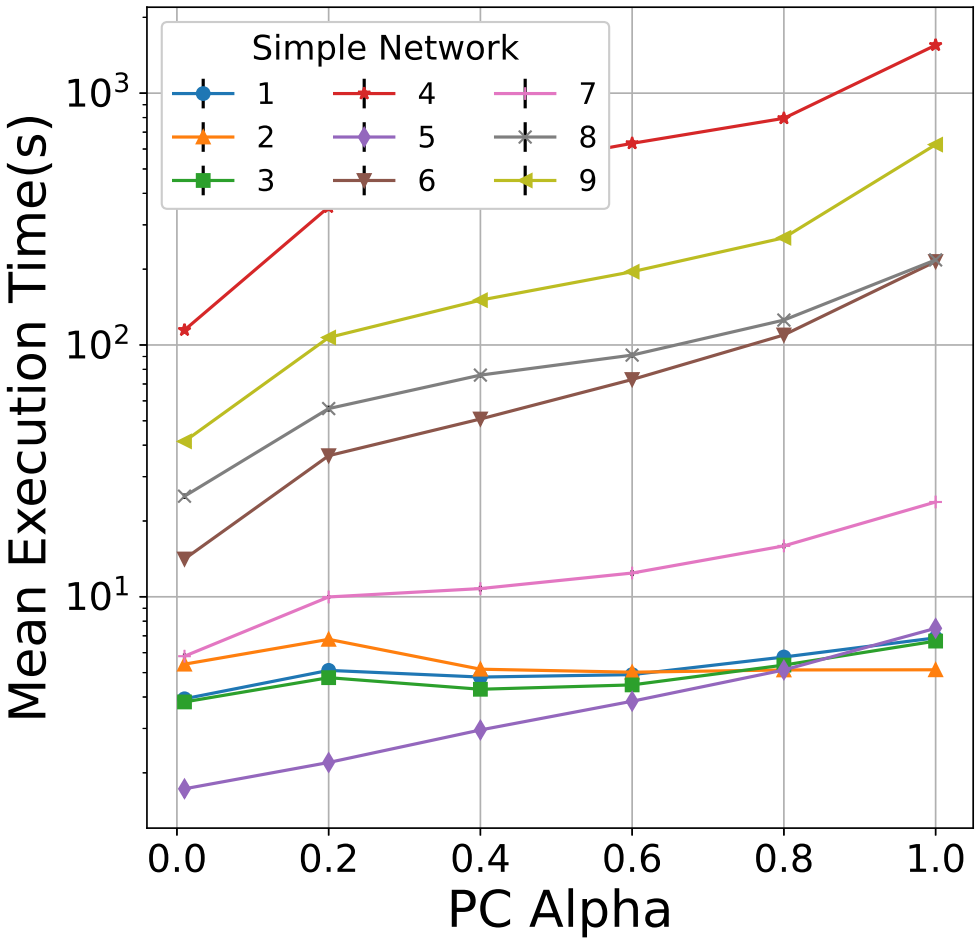
Supplementary Figure 12: Detailed performance curves of PCMCI$^+$ over simulated fMRI from simple networks for varying values of its hyperparameter PC Alpha. In all graphs, the error bars depict the standard error of the mean.



Supplementary Figure 13: Execution run times of FASK over simulated fMRI from simple networks for varying values of its hyperparameter Alpha. Error bars depict the standard error of the mean.

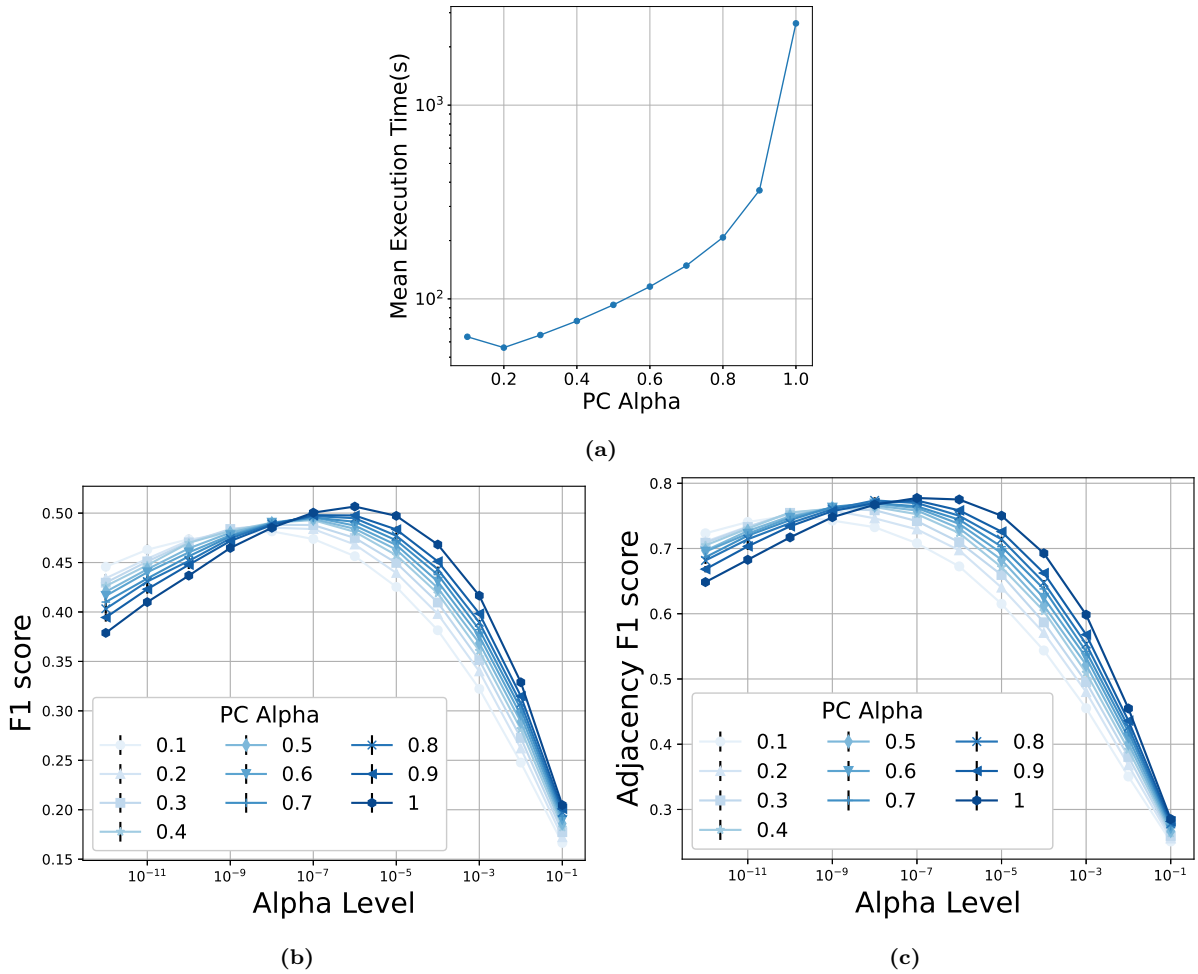


Supplementary Figure 14: Execution run times of DGlearn over simulated fMRI from simple networks for varying values of its hyperparameter BIC coefficient. In all graphs, the error bars depict the standard error of the mean.



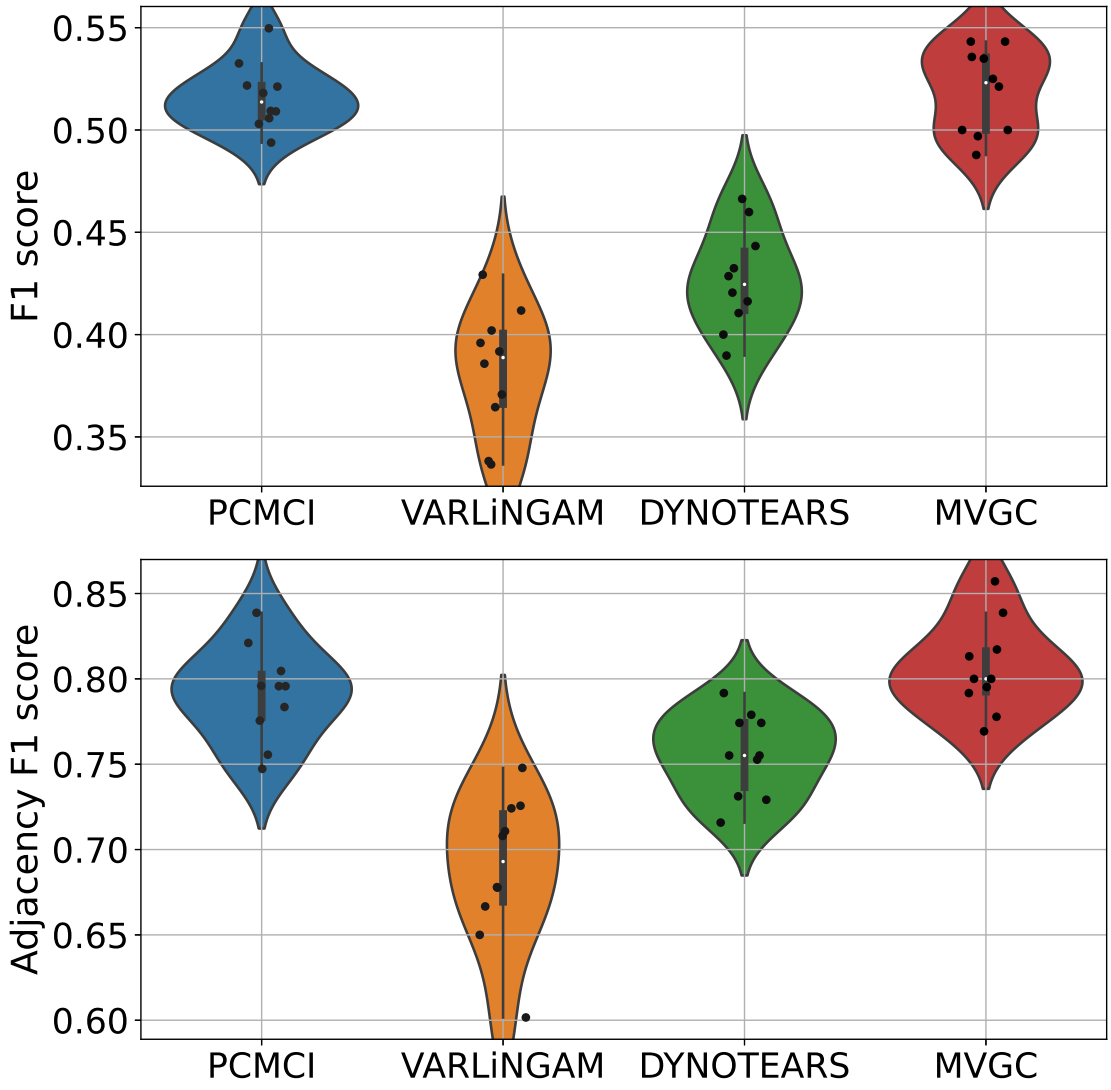
Supplementary Figure 15: Execution run times of PCMCI$^+$ over simulated fMRI from simple networks for varying values of its hyperparameter PC Alpha. Error bars depict the standard error of the mean.

Supplementary Figures for Simulated fMRI from Small-Degree Macaque Network

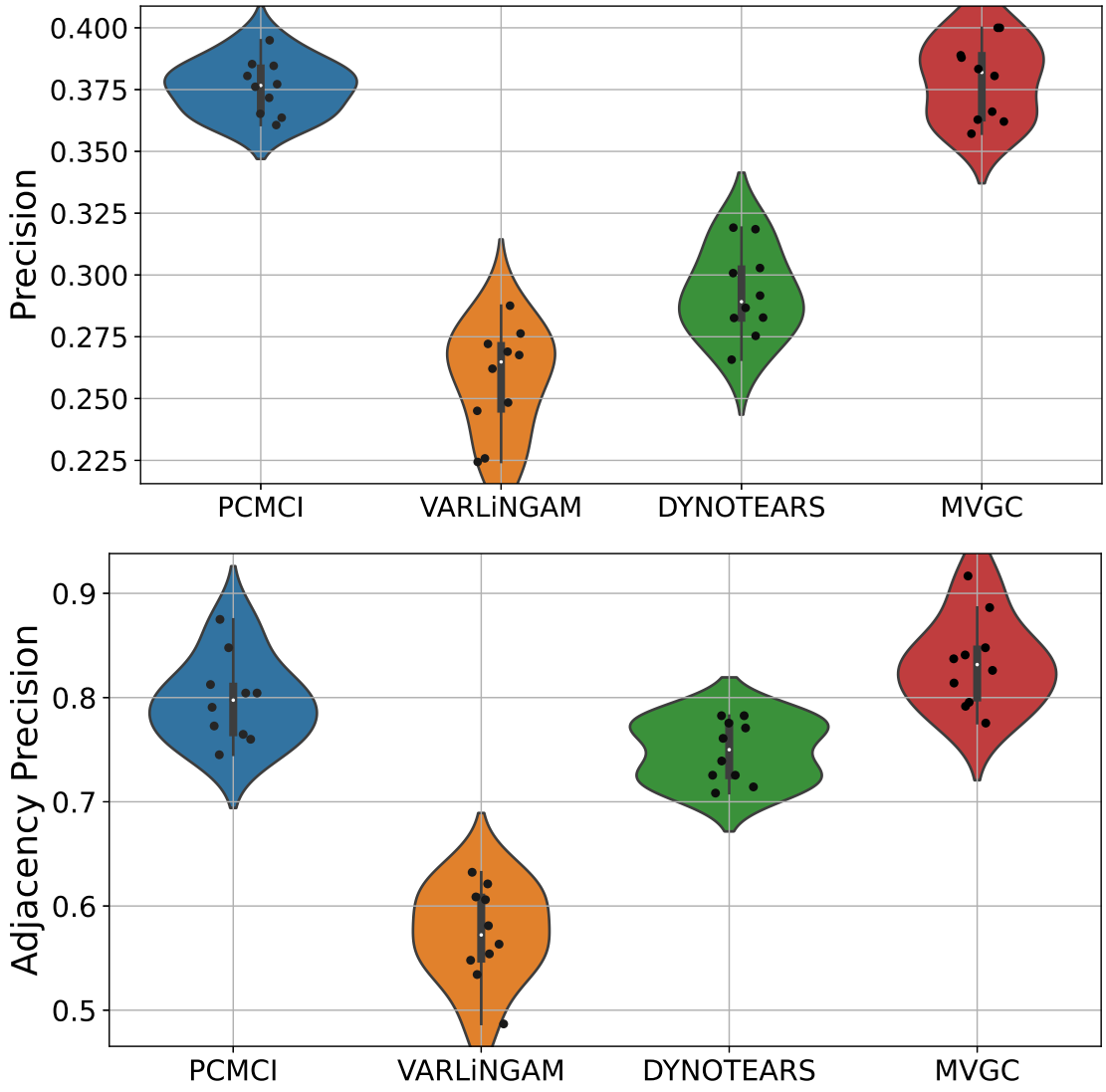


(a) **(b)** **(c)**

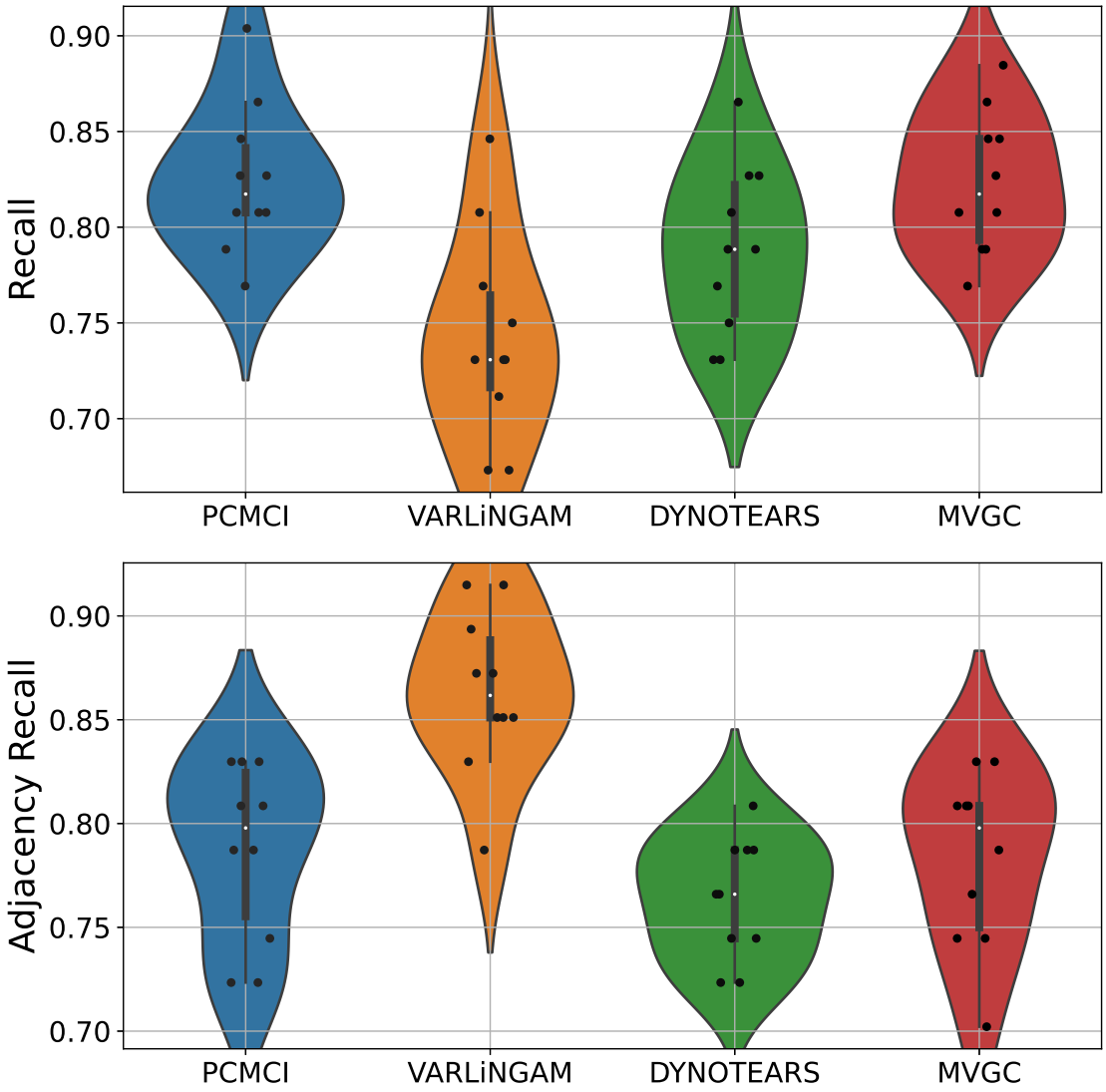
Supplementary Figure 16: The effect of the 'PC Alpha' hyperparameter of PCMCI on its performance over simulated Small-Degree Macaque fMRI. (a) Mean Execution time for different values of PC Alpha, showing a super-exponential growth. (b,c) F1 scores for varying values of PC Alpha and Alpha Level. Alpha Level denotes a second hyperparameter that controls the sparsity of the returned graph (the higher the Alpha Level the denser the resulting graphs). Best F1 scores are achieved for PC Alpha = 1. In all graphs, the error bars depict the standard error of the mean.



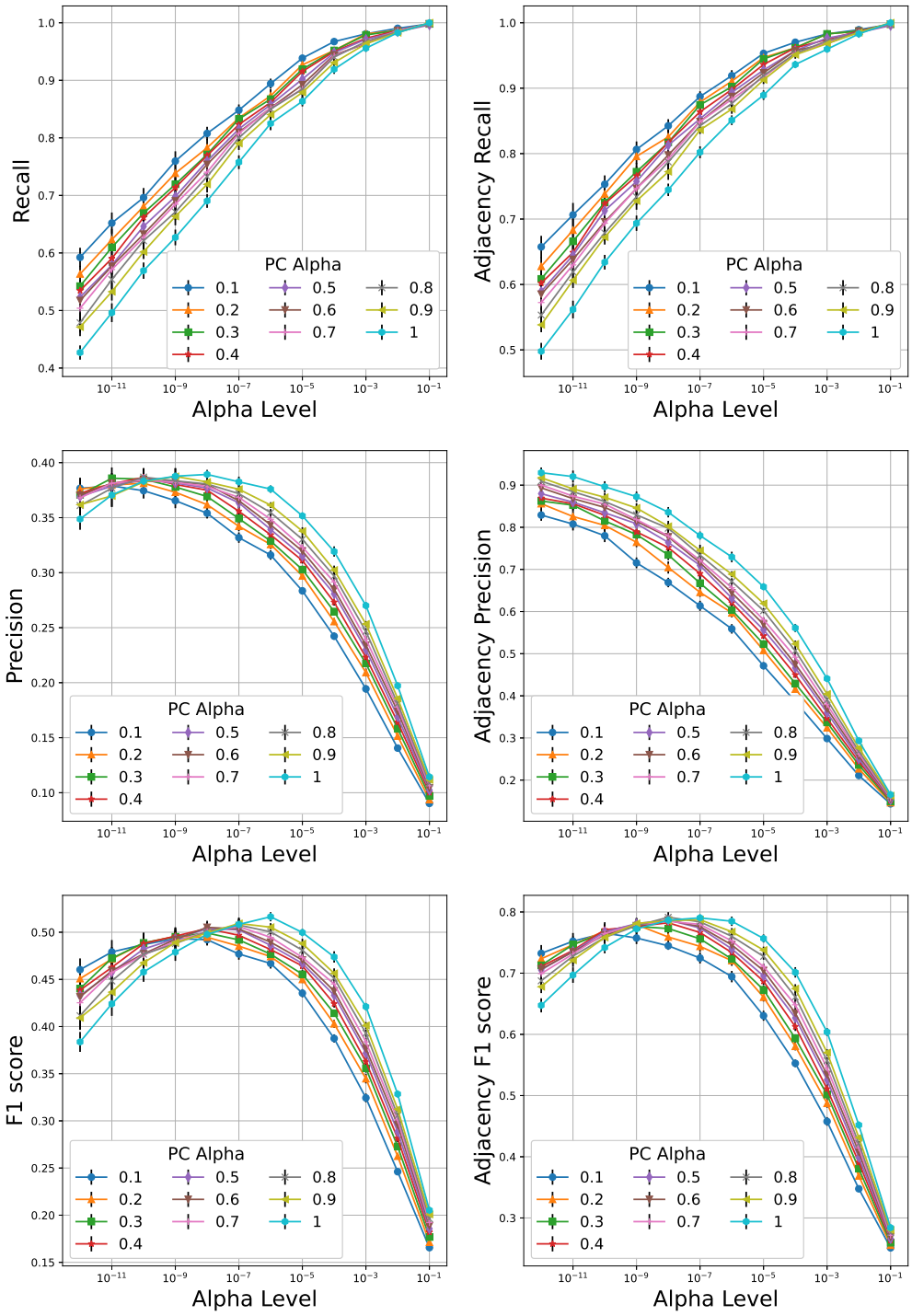
Supplementary Figure 17: Comparing the scalable subset of algorithms from Figure 1 over simulated fMRI data from the Small Degree Macaque benchmark . F1 score of identifying the full (directed) graph (top). Each distribution is calculated based on 10 repetitions of simulated data from the same underlying graph. Corresponding Adjacency F1 score of identifying the undirected skeleton of the Small Degree Macaque network (bottom).



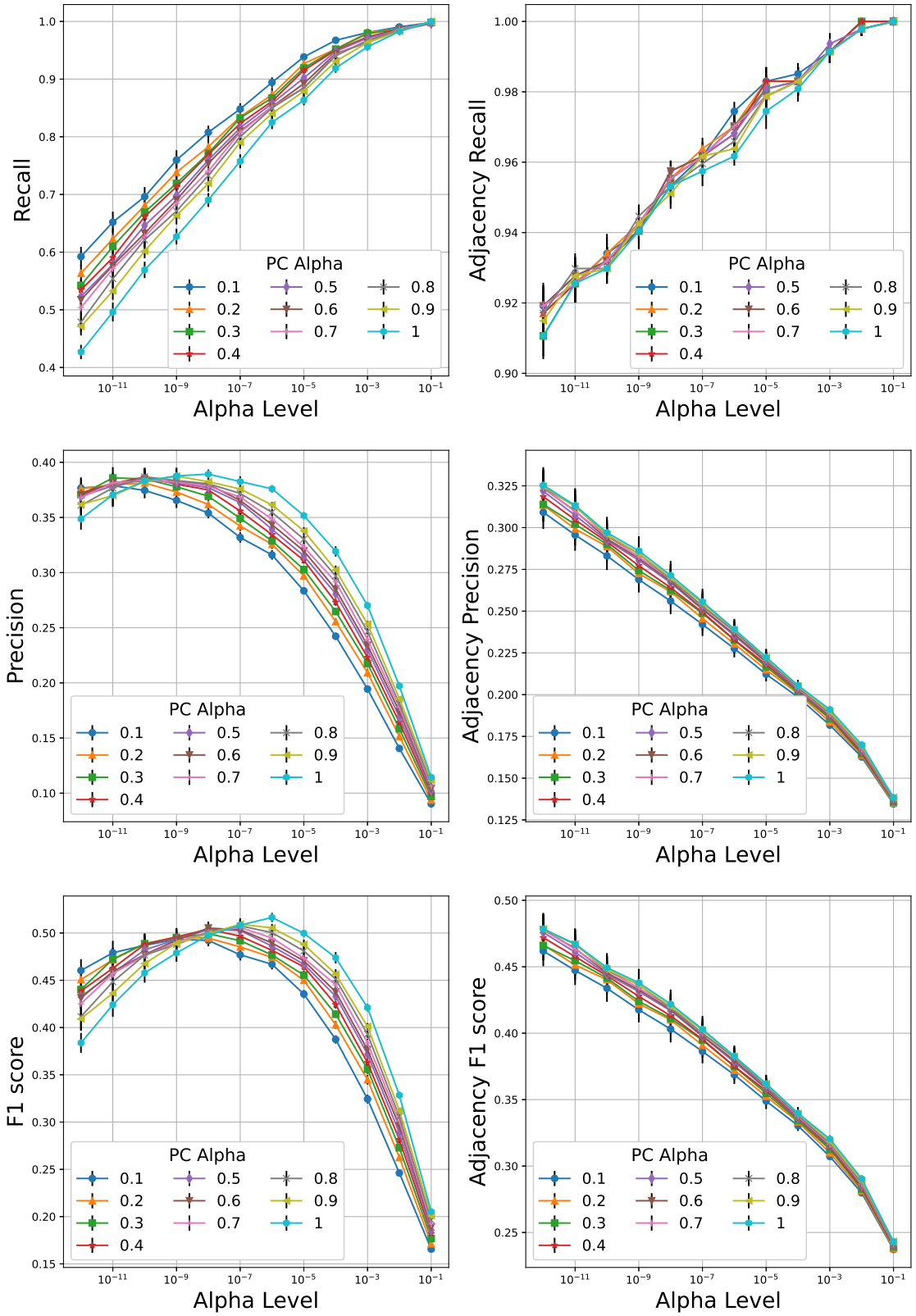
Supplementary Figure 18: Comparing the scalable subset of algorithms from Figure 1 over simulated fMRI data from the Small Degree Macaque benchmark . Precision of identifying the full (directed) graph (top). Each distribution is calculated based on 10 repetitions of simulated data from the same underlying graph. Corresponding Adjacency precision of identifying the undirected skeleton of the Small Degree Macaque network (bottom).



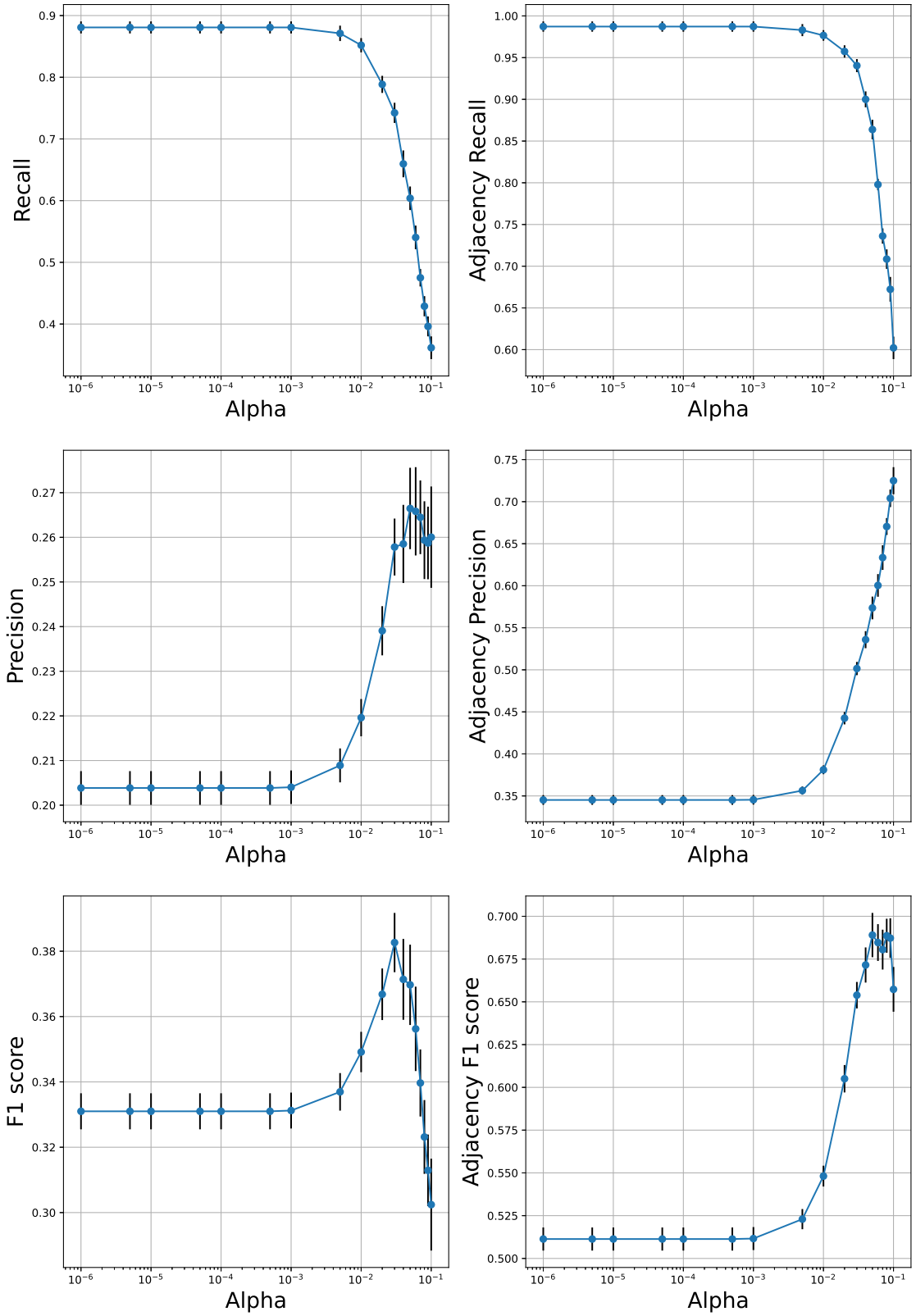
Supplementary Figure 19: Comparing the scalable subset of algorithms from Figure 1 over simulated fMRI data from the Small Degree Macaque benchmark . Recall of identifying the full (directed) graph (top). Each distribution is calculated based on 10 repetitions of simulated data from the same underlying graph. Corresponding Adjacency recall of identifying the undirected skeleton of the Small Degree Macaque network (bottom).



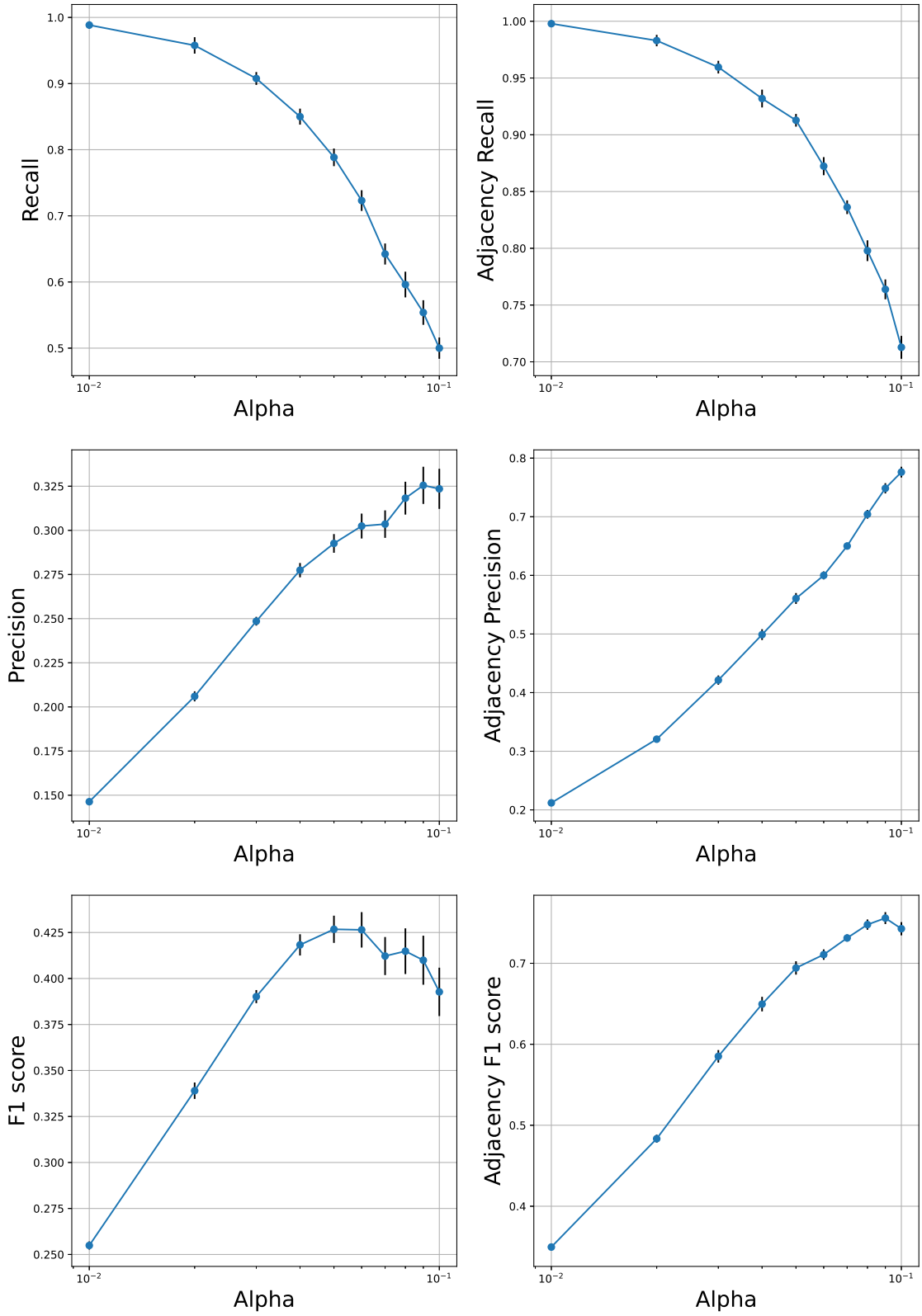
Supplementary Figure 20: Detailed performance curves of PCMCI over simulated fMRI from Small-Macaque network for varying values of its hyperparameters PC Alpha and Alpha Level. In all graphs, the error bars depict the standard error of the mean.



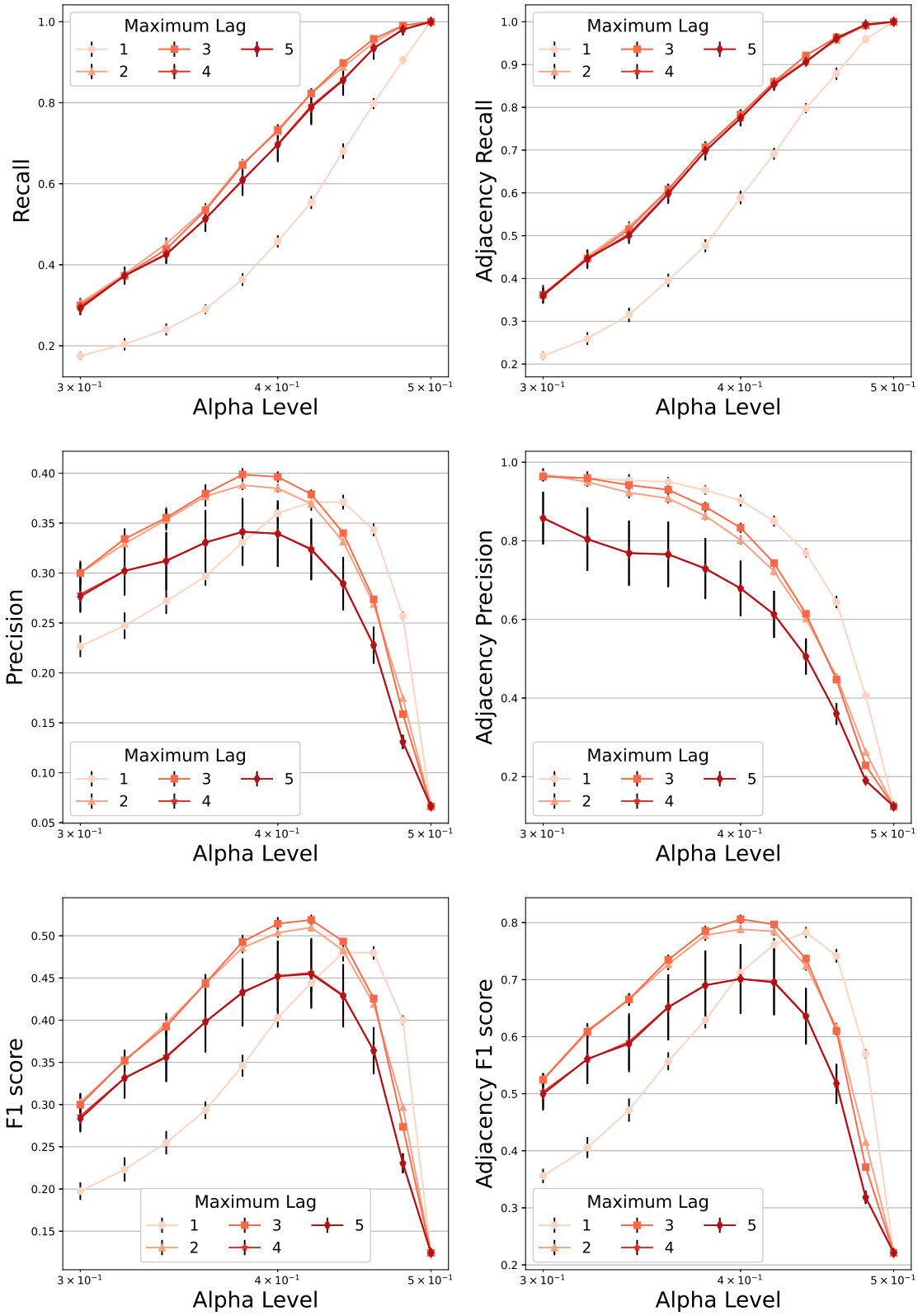
Supplementary Figure 21: Detailed performance curves of Mixed-PCMCI over simulated fMRI from Small-Macaque network for varying values of its hyperparameters PC Alpha and Alpha Level. In all graphs, the error bars depict the standard error of the mean.



Supplementary Figure 22: Detailed performance curves of VARLiNGAM over simulated fMRI from Small-Macaque network for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.

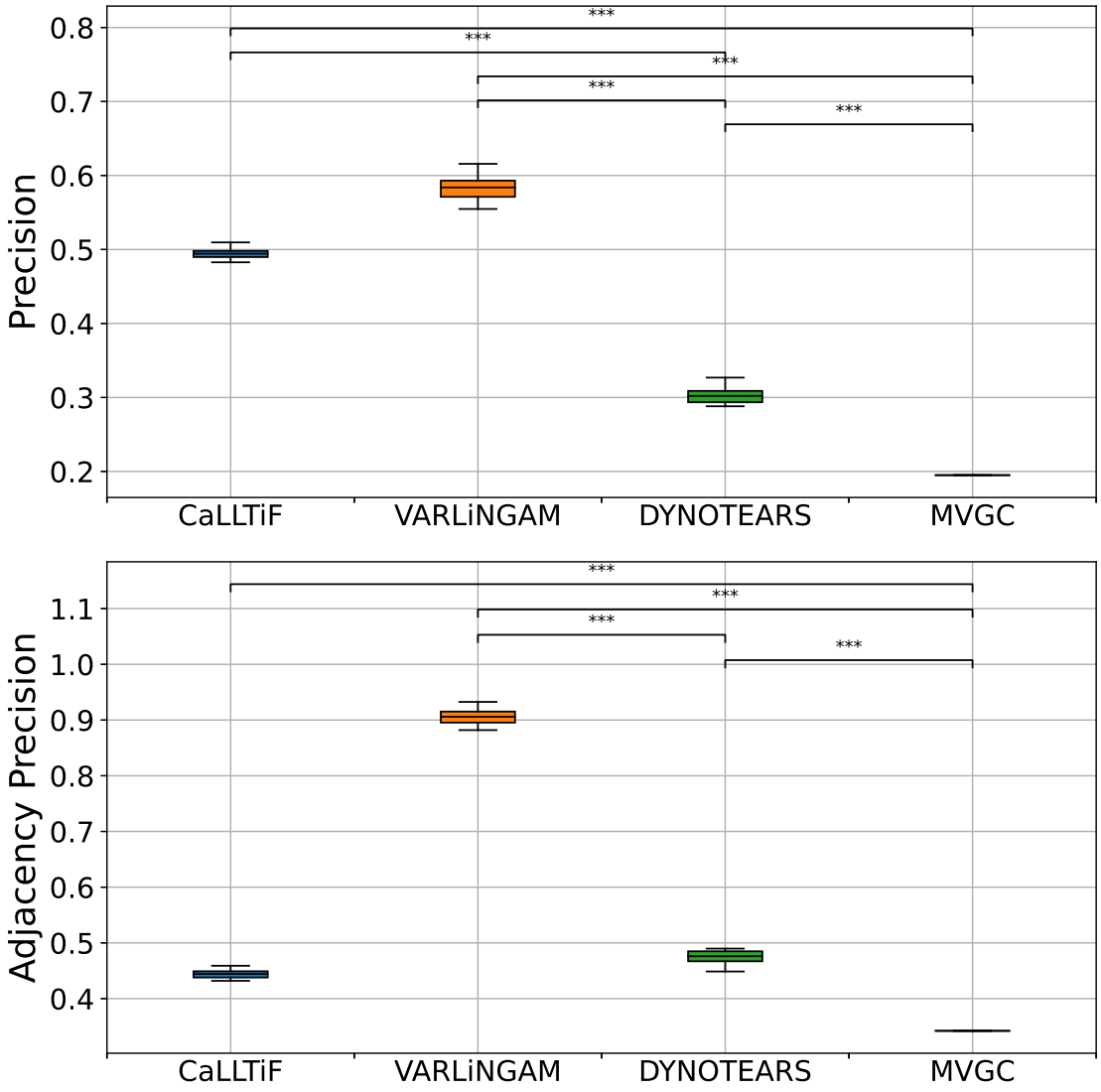


Supplementary Figure 23: Detailed performance curves of DYNOTEARS over simulated fMRI from Small-Macaque network for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.

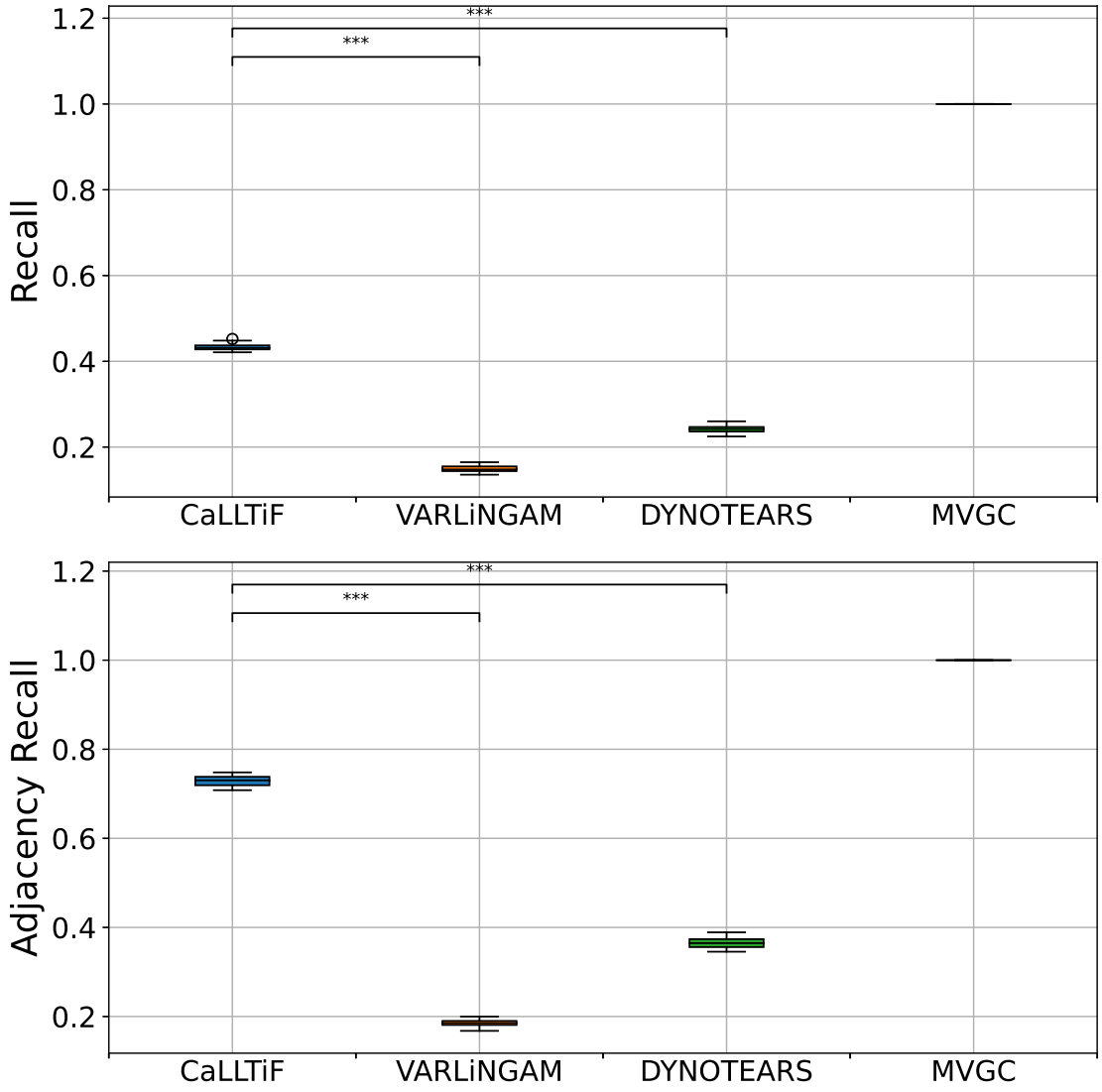


Supplementary Figure 24: Detailed performance curves of MVGC over simulated fMRI from Small-Macaque network for varying values of its hyperparameters Alpha and maximum number of lags. In all graphs, the error bars depict the standard error of the mean.

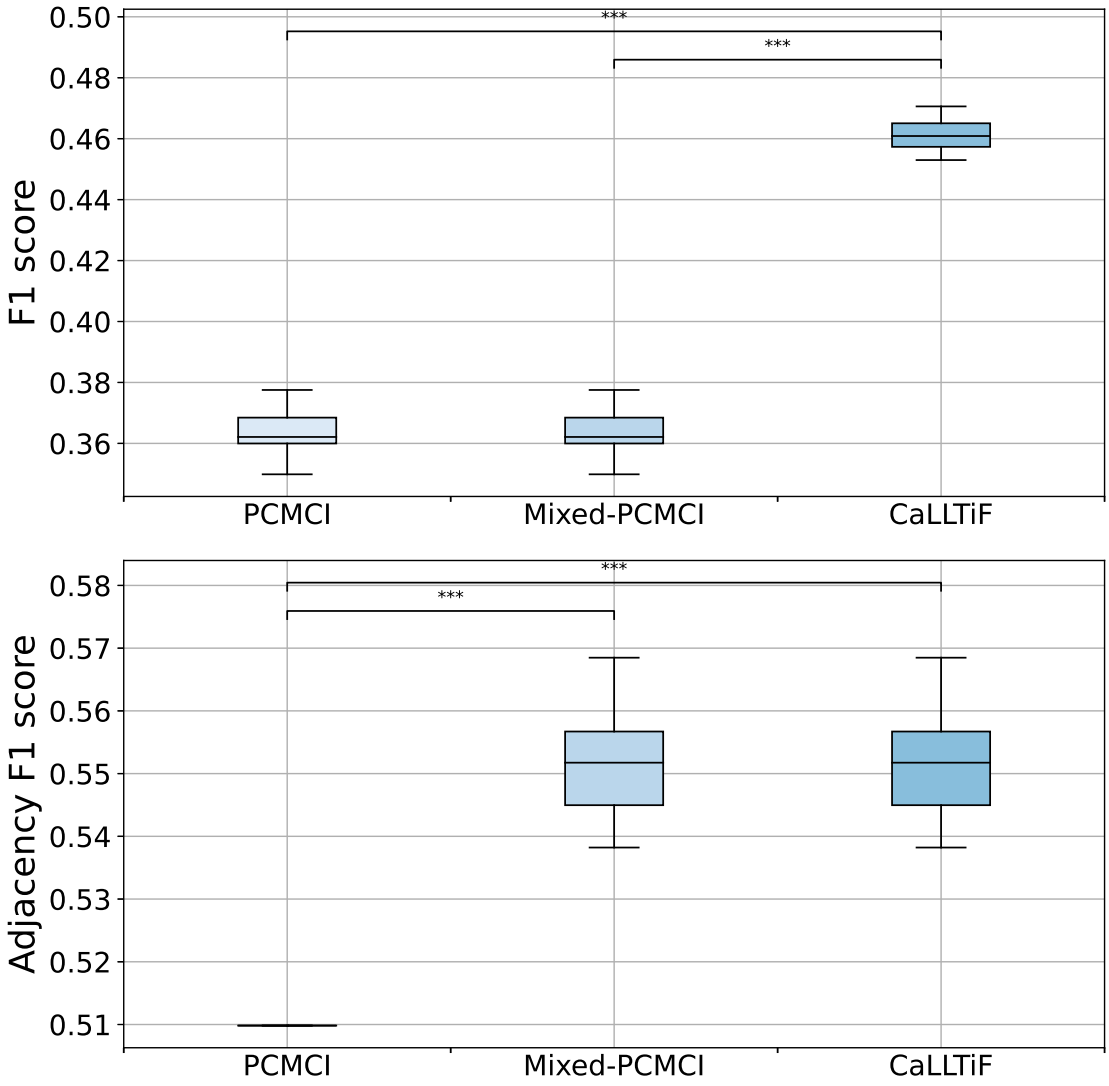
Supplementary Figures for Simulated fMRI from Full Macaque Network



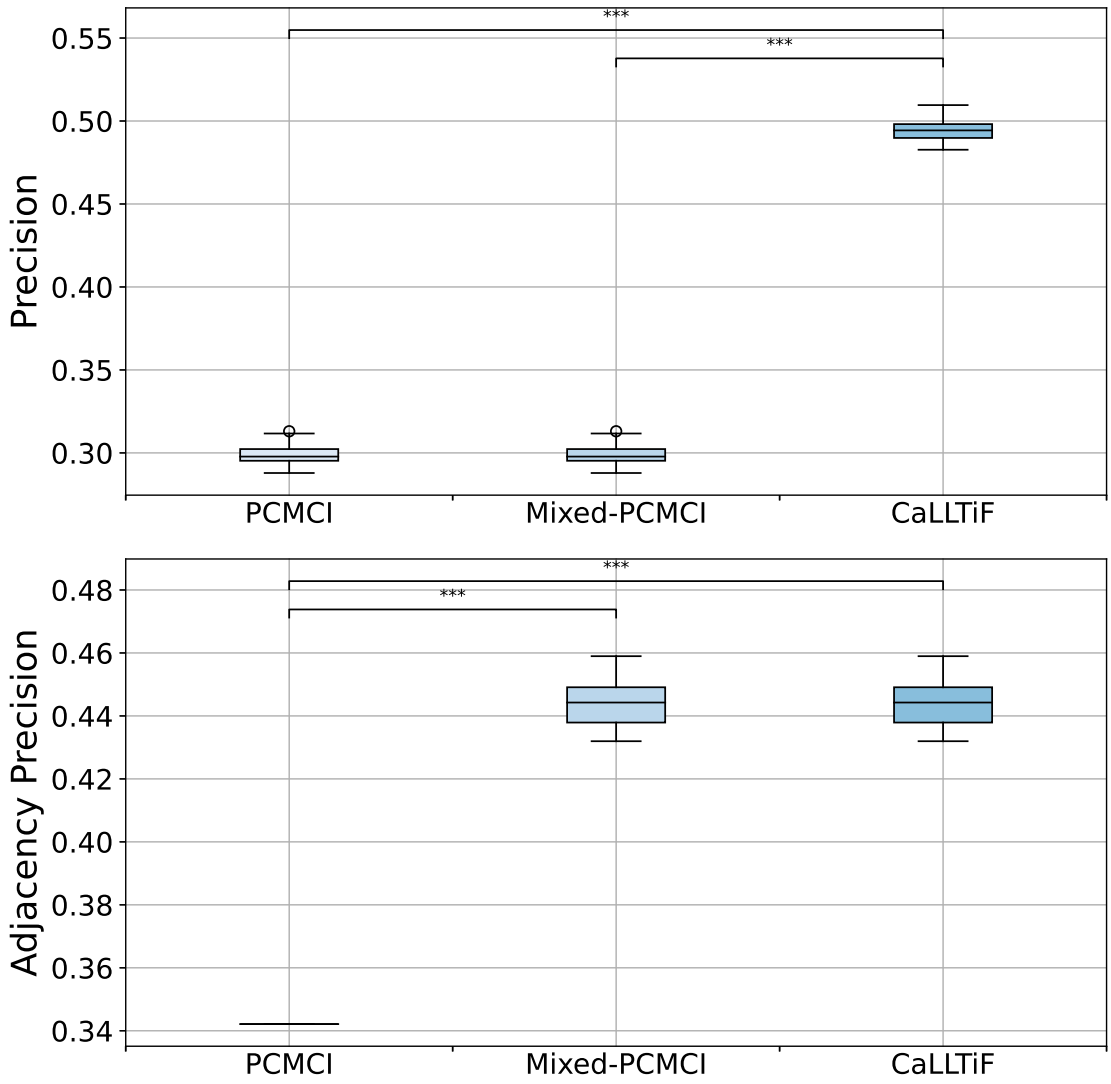
Supplementary Figure 25: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. Distributions of Precision and Recall for CaLLTiF and state-of-the-art alternatives. Precision is shown for the values of hyperparameters which give the maximum F1 score. For all repetitions, the best performance of MVGC occurs at $\alpha = 0.5$ which returns a complete graph, hence the point distributions for MVGC. *** denotes $p < 0.001$. All statistical comparisons are performed using a one-sided Wilcoxon signed-rank test. In all boxplots, the center line represents the median, the box spans the interquartile range (IQR), and the whiskers extend up to 1.5 times the IQR from the box limits.



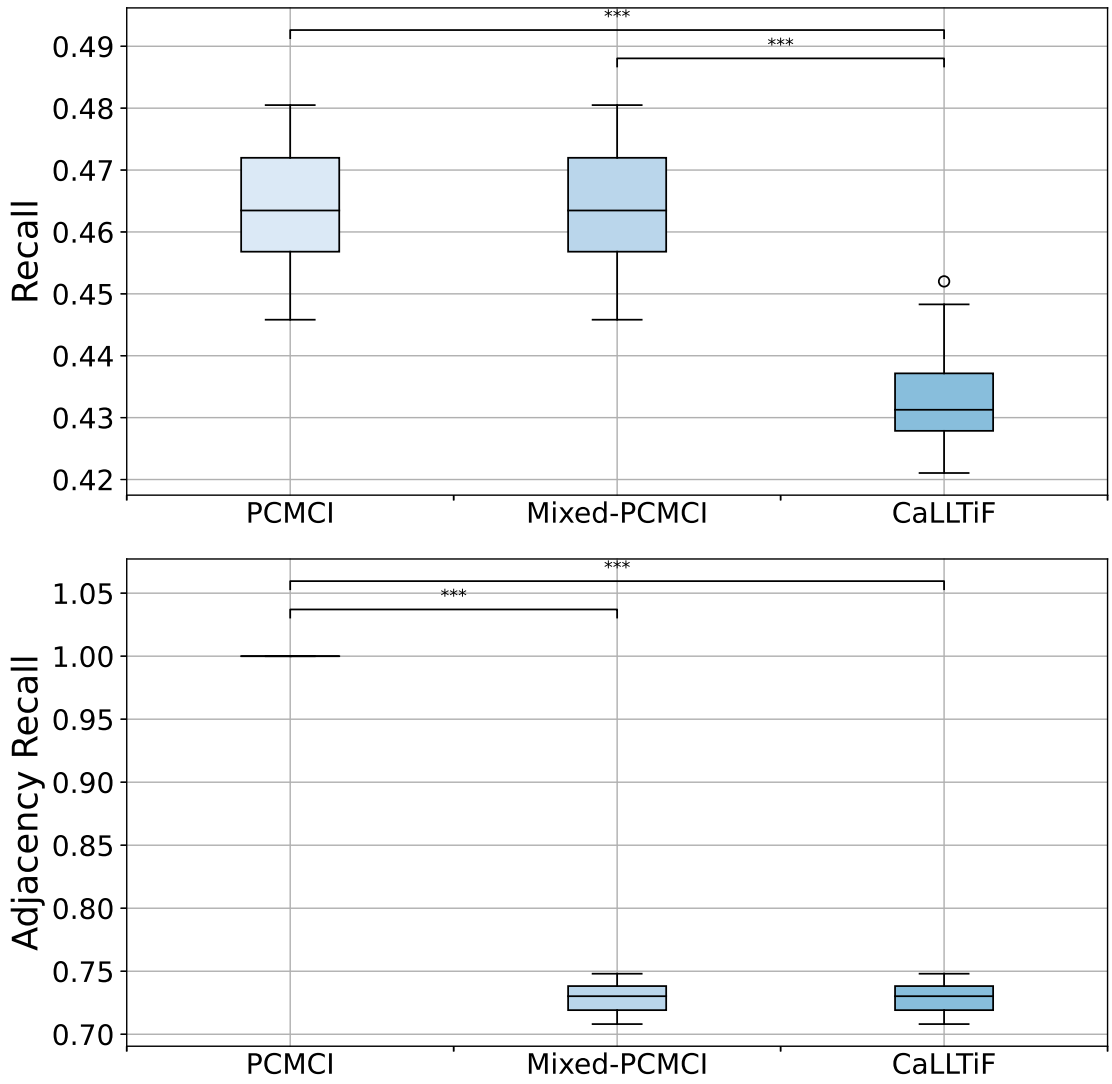
Supplementary Figure 26: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. Distributions of Precision and Recall for CaLLTiF and state-of-the-art alternatives. Recall is shown for the values of hyperparameters which give the maximum F1 score. For all repetitions, the best performance of MVGC occurs at $\alpha = 0.5$ which returns a complete graph, hence the point distributions for MVGC. *** denotes $p < 0.001$. All statistical comparisons are performed using a one-sided Wilcoxon signed-rank test. In all boxplots, the center line represents the median, the box spans the interquartile range (IQR), and the whiskers extend up to 1.5 times the IQR from the box limits.



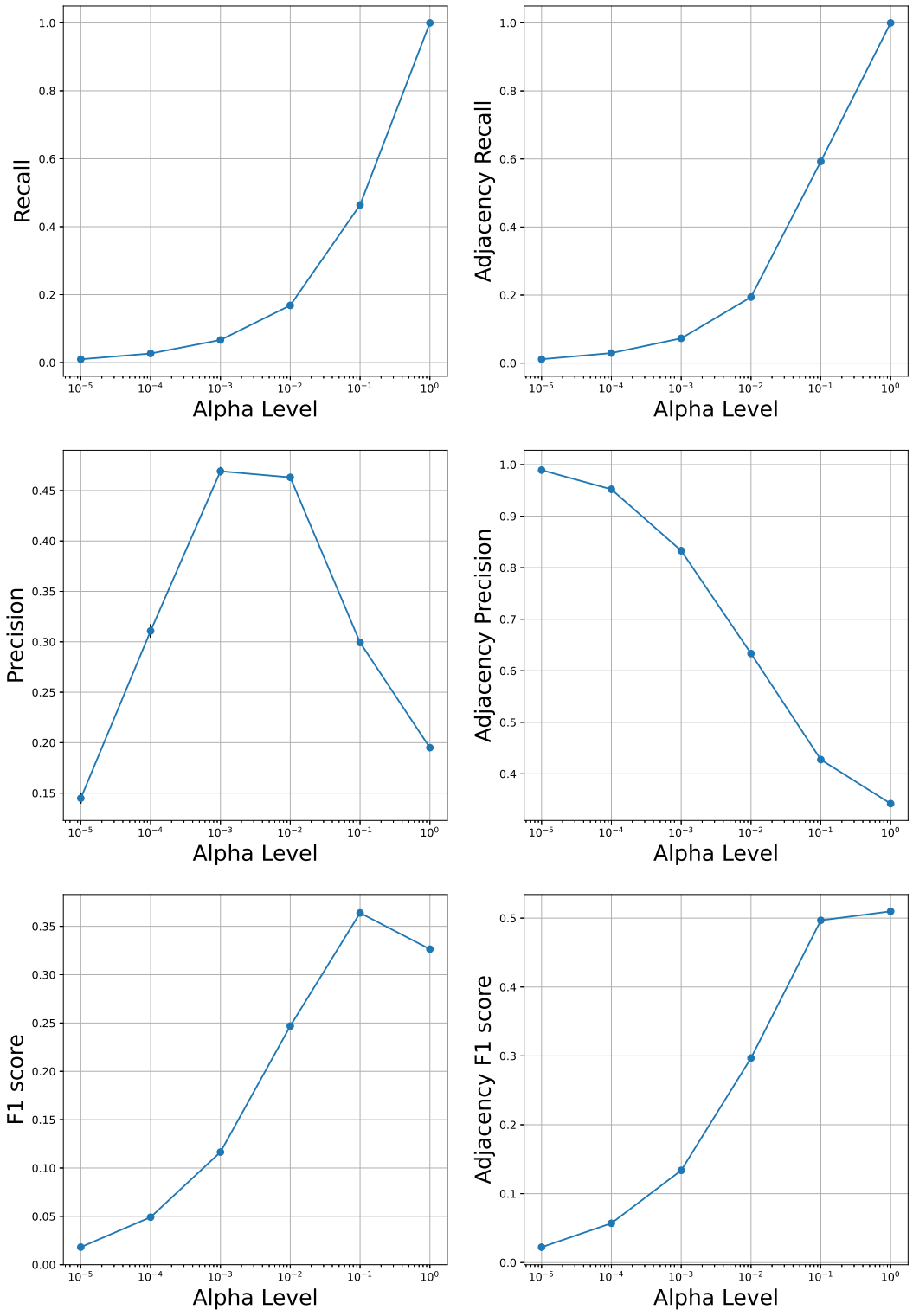
Supplementary Figure 27: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. Distributions of F1 score for PCMCI (ignoring the contemporaneous ⊶⊷ connections), Mixed-PCMCI (using the contemporaneous ⊶⊷ connections only for adjacency), and CaLLTiF. F1 scores are shown for the value of hyperparameters which give the maximum F1 score.



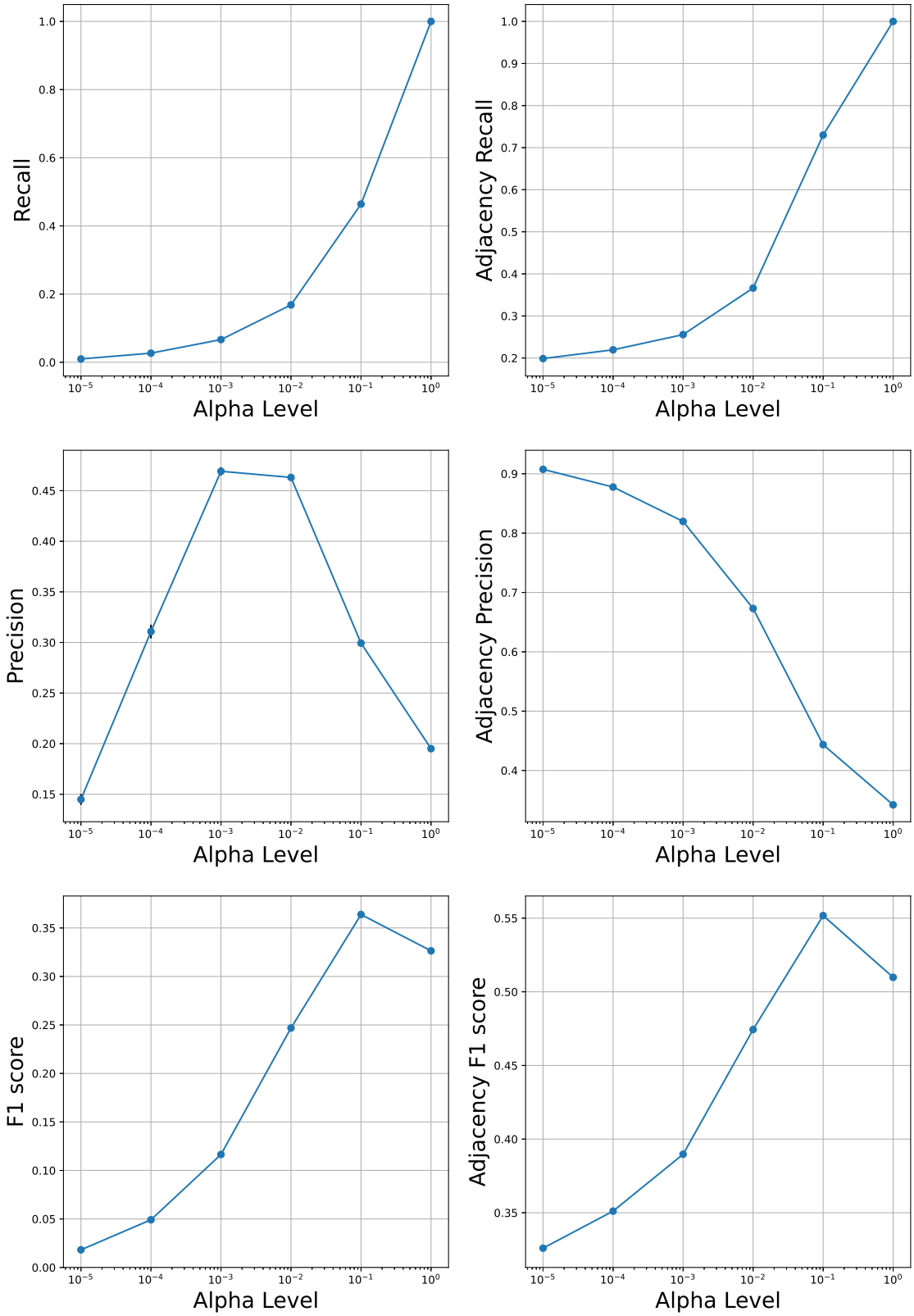
Supplementary Figure 28: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. Distributions of precision for PCMCI (ignoring the contemporaneous ⊶ connections), Mixed-PCMCI (using the contemporaneous ⊶ connections only for adjacency), and CaLLTiF. Precisions are shown for the value of hyperparameters which give the maximum F1 score.



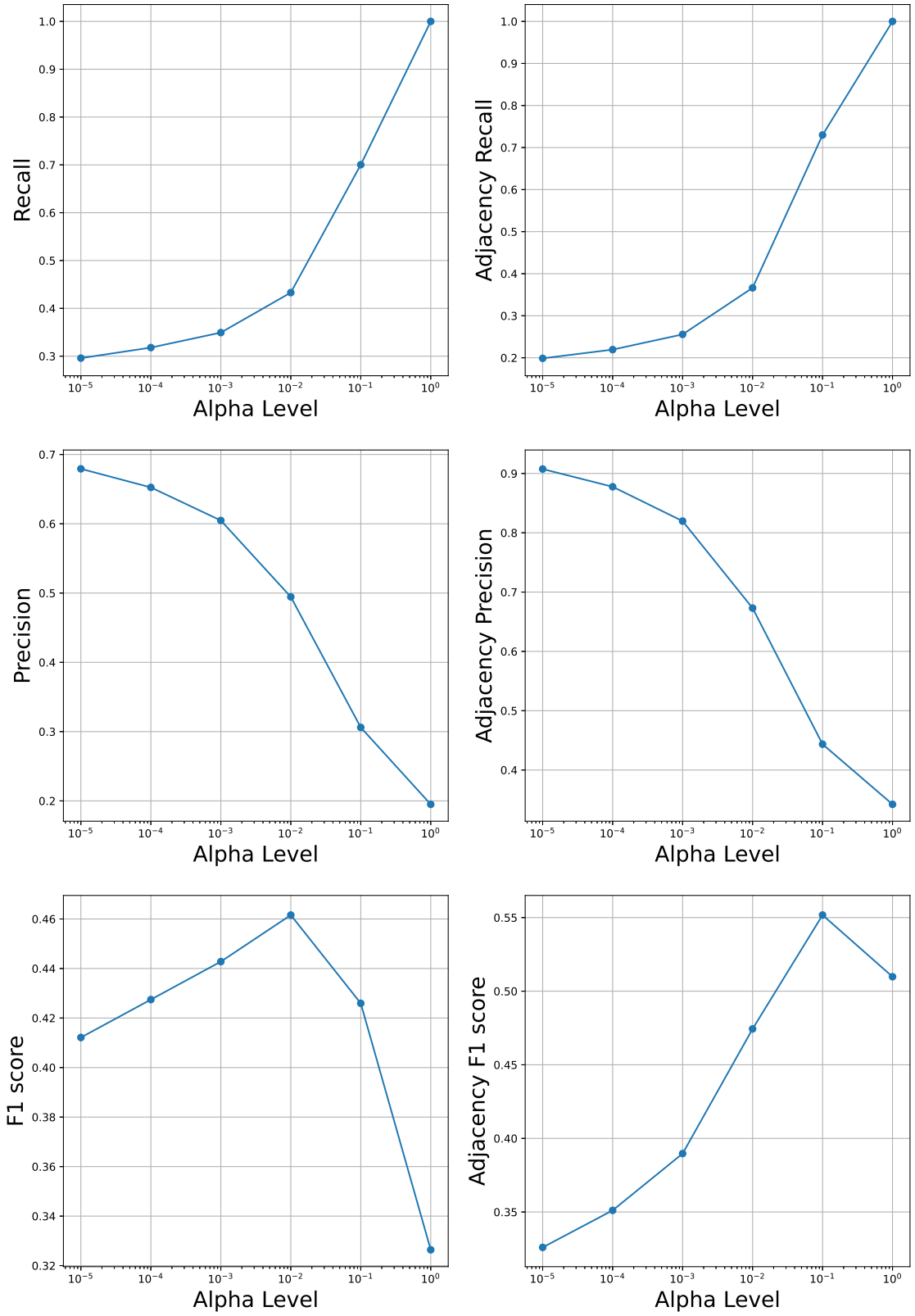
Supplementary Figure 29: Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome. Distributions of recall for PCMCI (ignoring the contemporaneous ⊶ connections), Mixed-PCMCI (using the contemporaneous ⊶ connections only for adjacency), and CaLLTiF. Recalls are shown for the value of hyperparameters which give the maximum F1 score.



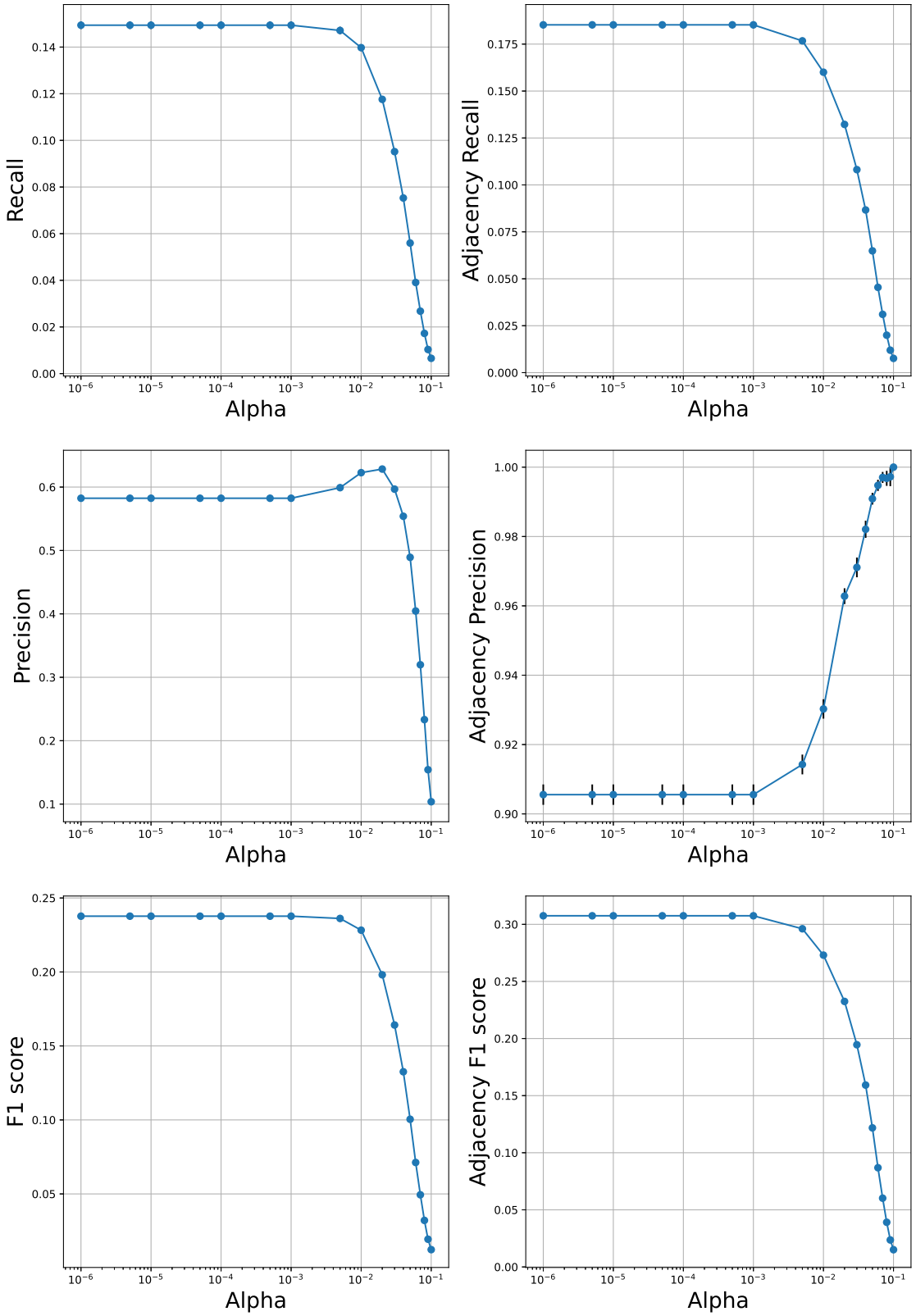
Supplementary Figure 30: Detailed performance curves of PCMCI over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha Level. In all graphs, the error bars depict the standard error of the mean.



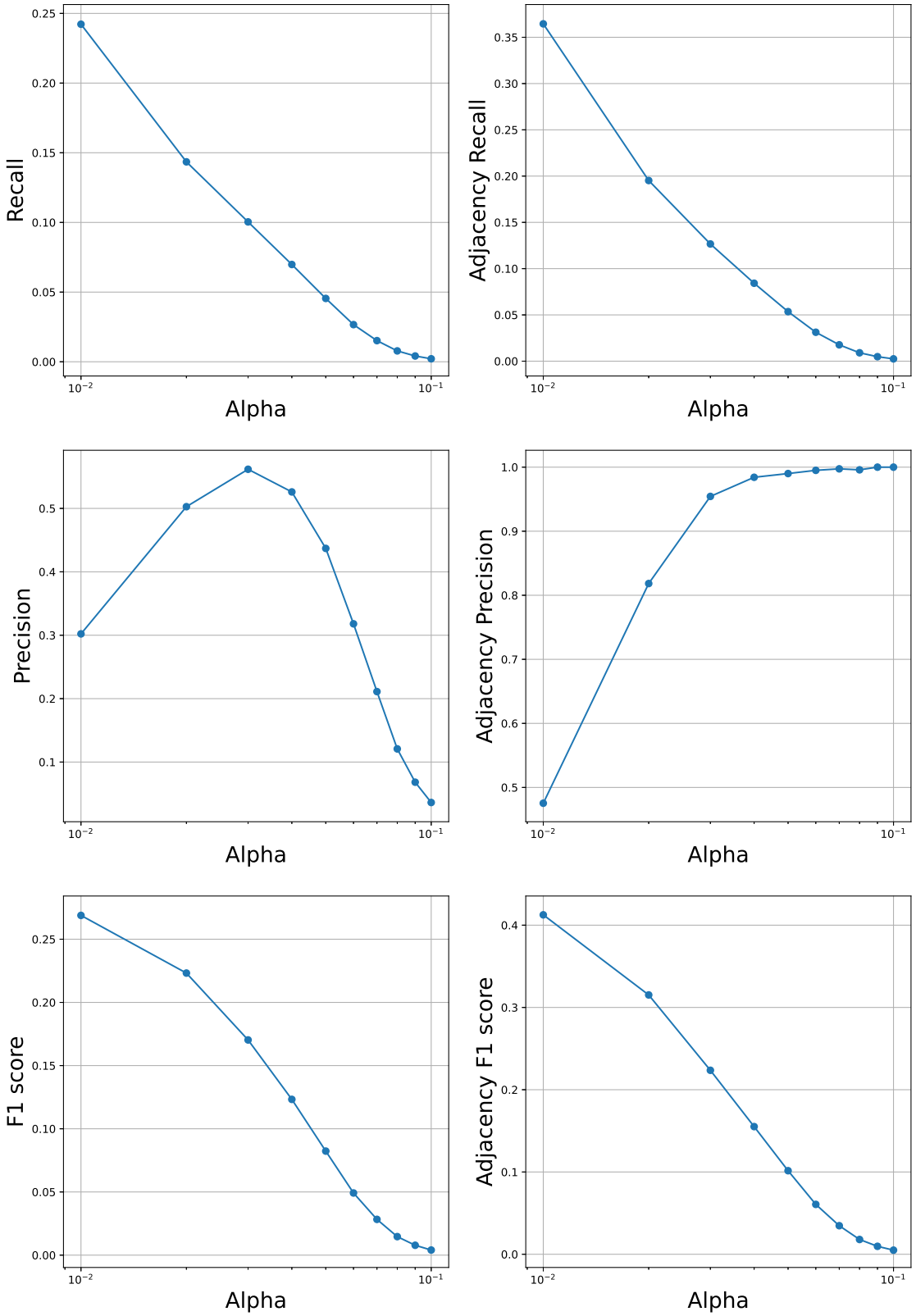
Supplementary Figure 31: Detailed performance curves of Mixed-PCMCI over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha Level. In all graphs, the error bars depict the standard error of the mean.



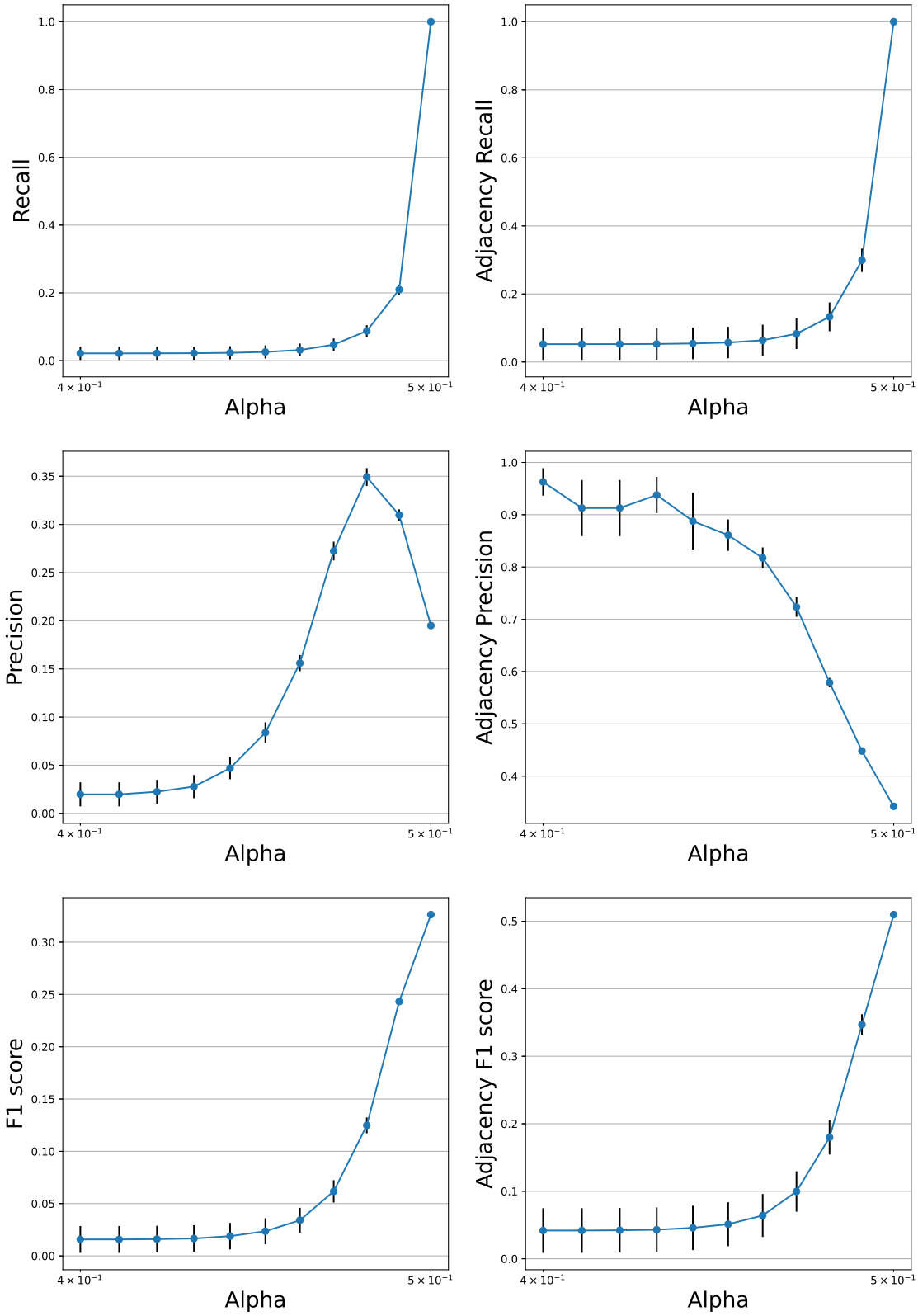
Supplementary Figure 32: Detailed performance curves of CaLLTiF over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha Level. In all graphs, the error bars depict the standard error of the mean.



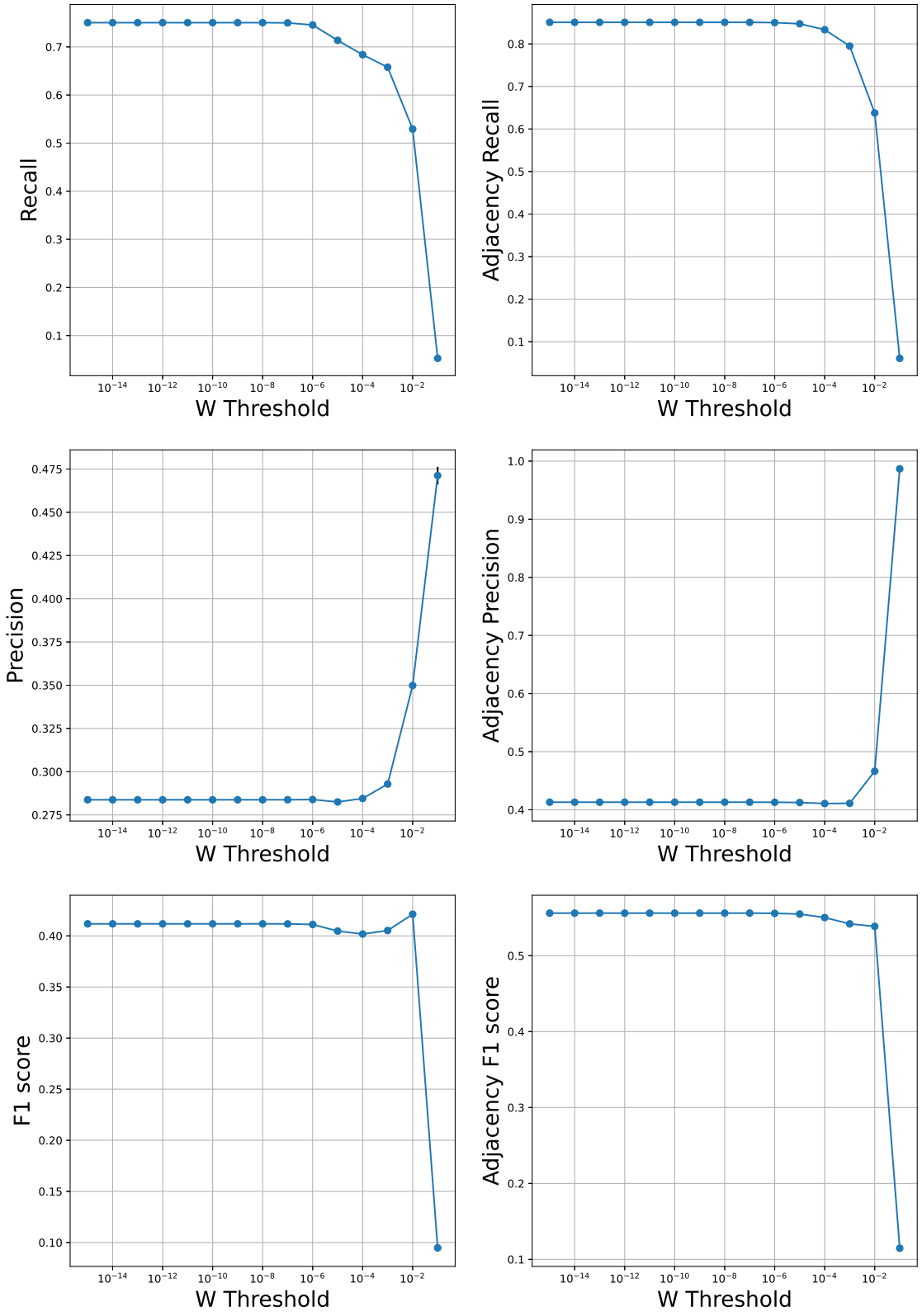
Supplementary Figure 33: Detailed performance curves of VARLiNGAM over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



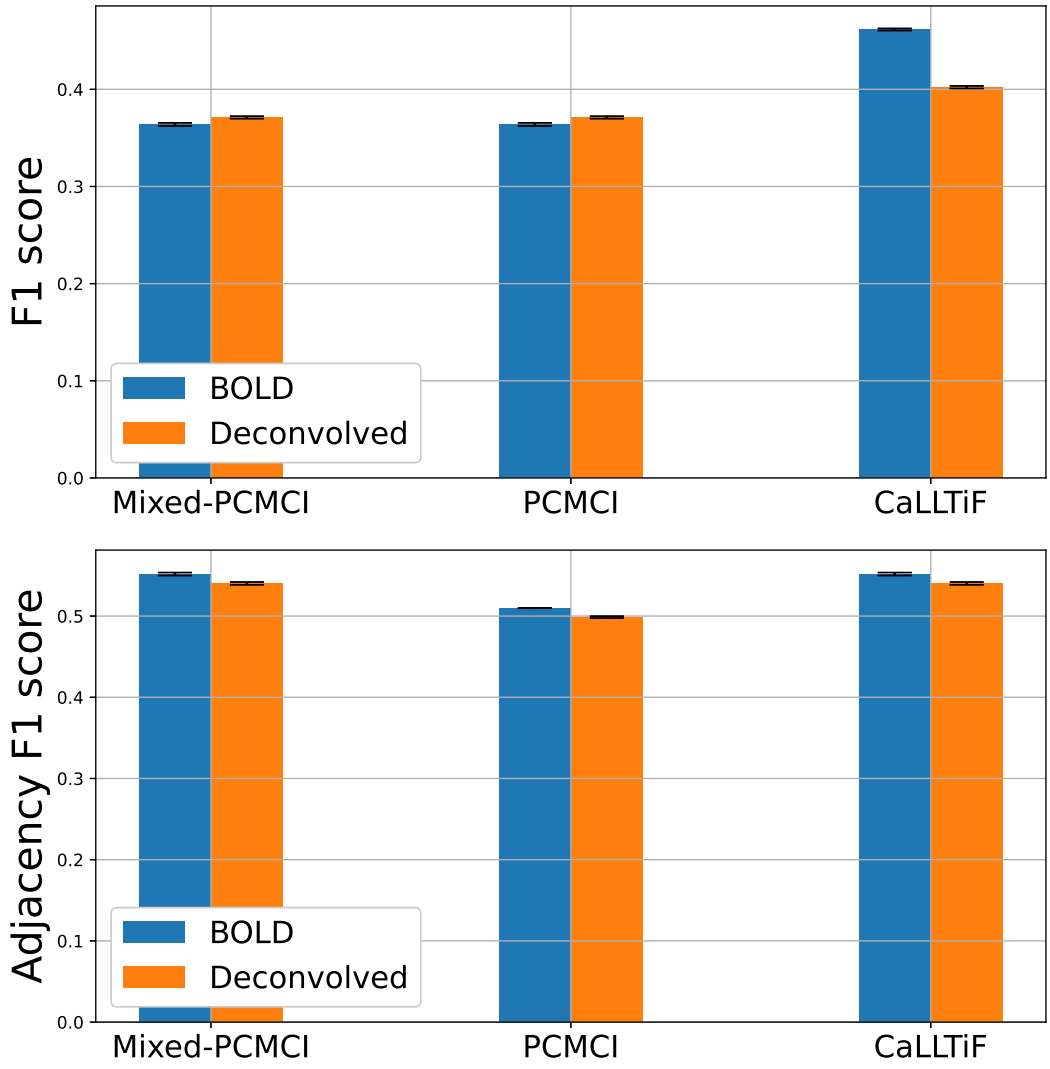
Supplementary Figure 34: Detailed performance curves of DYNOTEARS over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



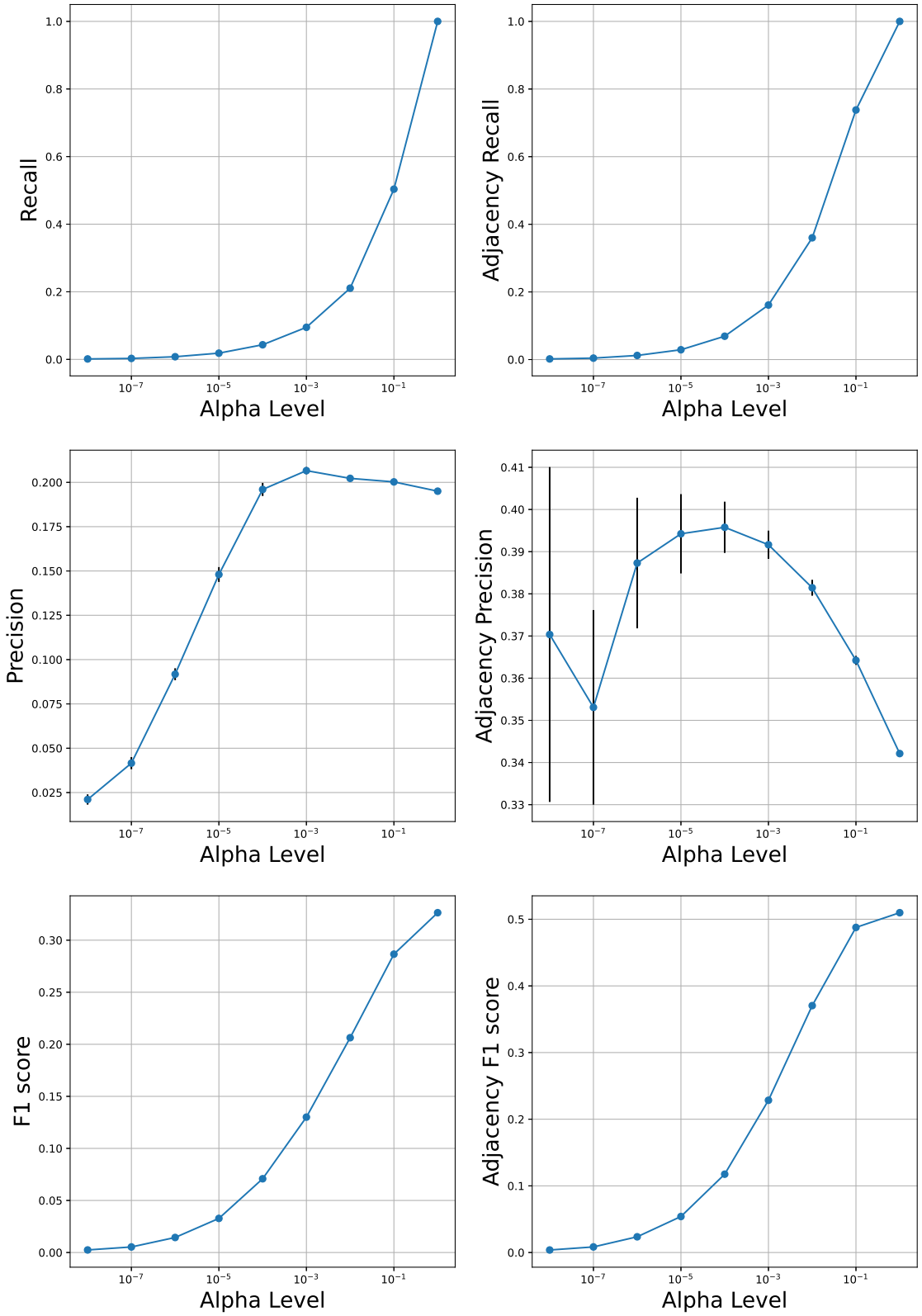
Supplementary Figure 35: Detailed performance curves of MVGC over simulated fMRI from Full Macaque network for varying values of its hyperparameter Alpha. In all graphs, the error bars depict the standard error of the mean.



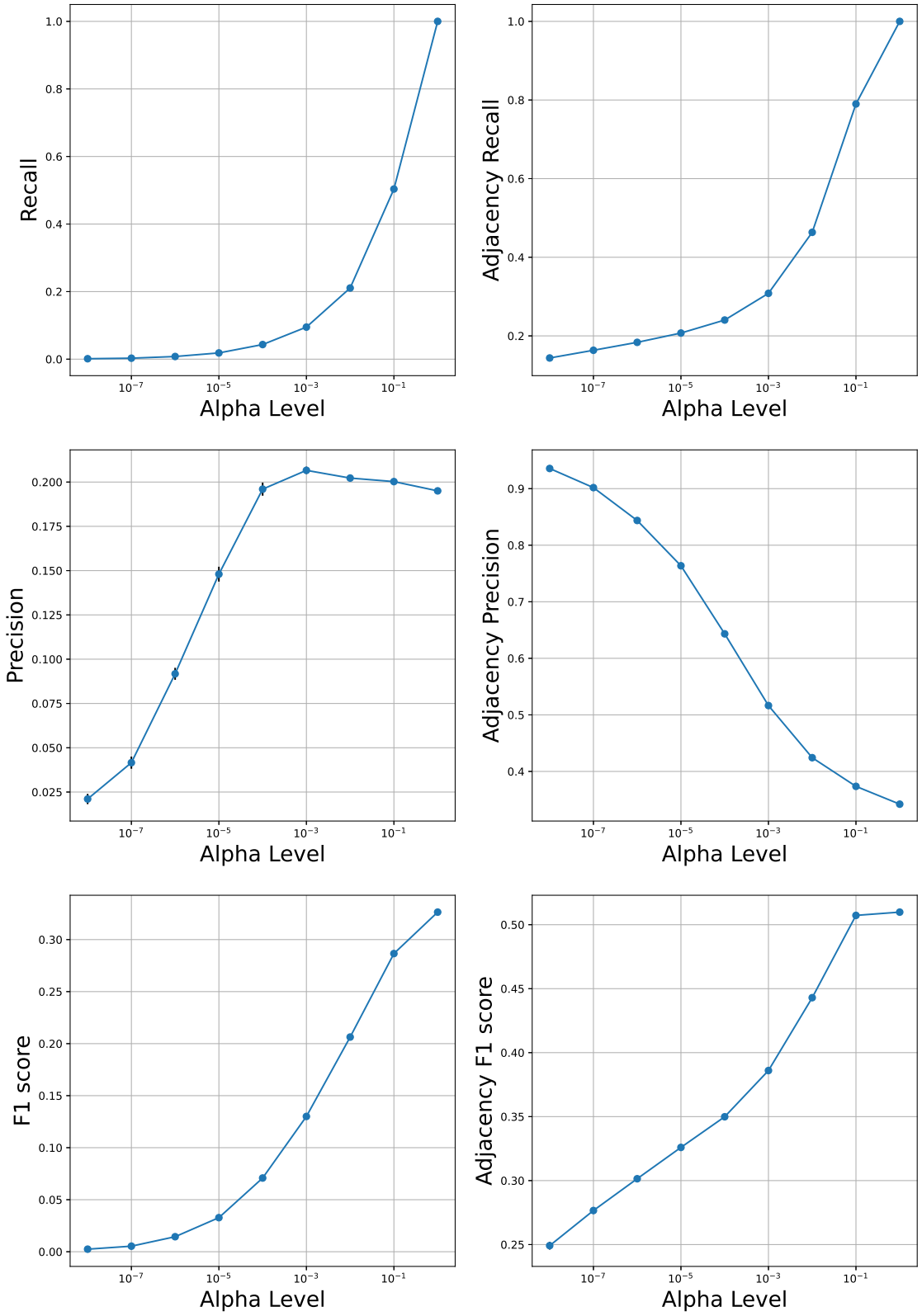
Supplementary Figure 36: Detailed performance curves of NTS-NOTEARS over simulated fMRI from Full Macaque network for varying values of its hyperparameter W Threshold. In all graphs, the error bars depict the standard error of the mean.



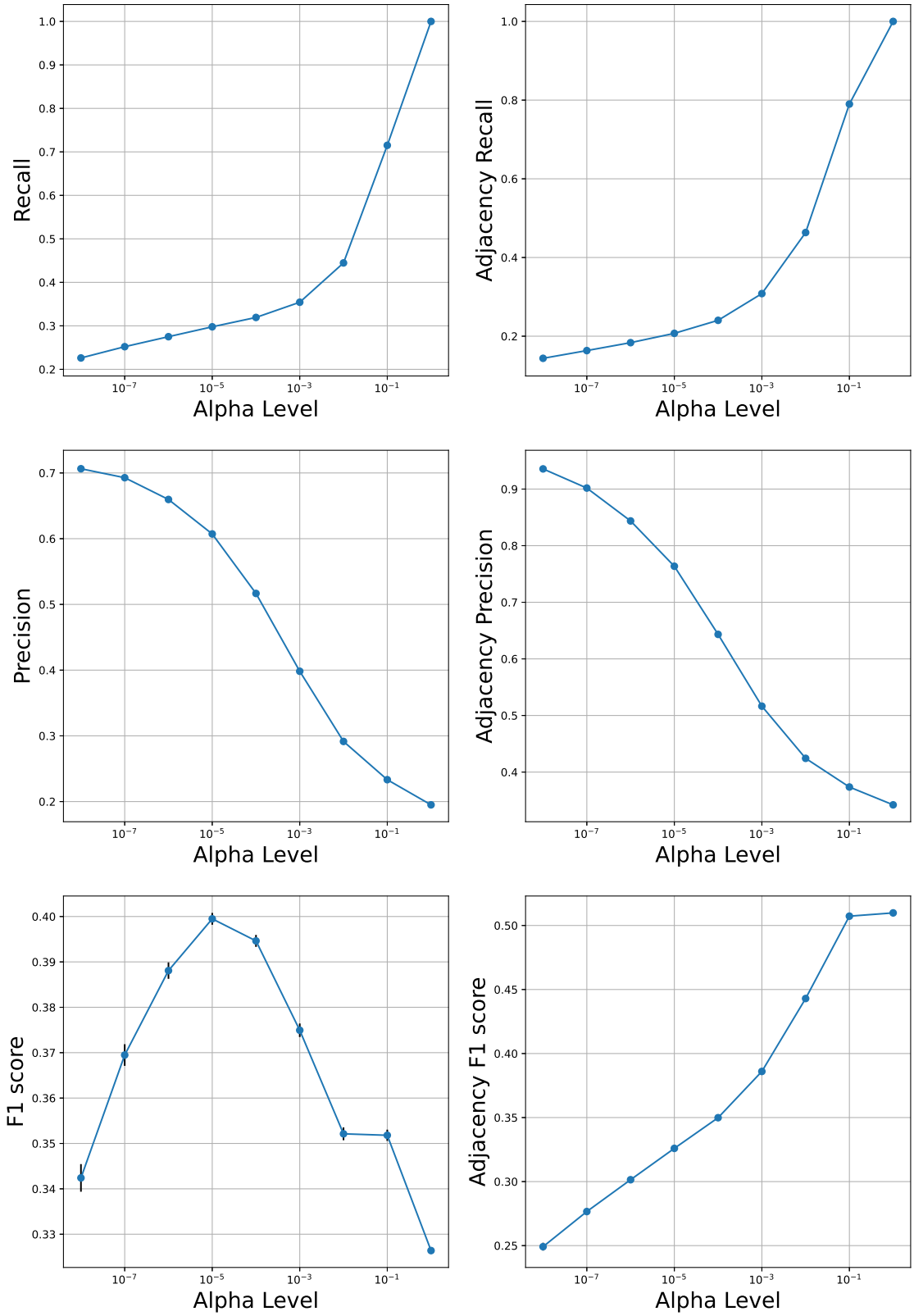
Supplementary Figure 37: Comparison between F1 scores and adjacency F1 scores of Mixed-PCMCI, PCMCI, and CaLLTiF when applied to simulated fMRI from the Full Macaque network with (orange, right bars) and without (blue, left bars) deconvolution with a generic hemodynamic response function (HRF). In all graphs, the error bars depict one standard deviation.



Supplementary Figure 38: Detailed performance curves of PCMCI over simulated fMRI from Full Macaque network when using ground-truth parent sets for each conditional independence test. In all graphs, the error bars depict the standard error of the mean.

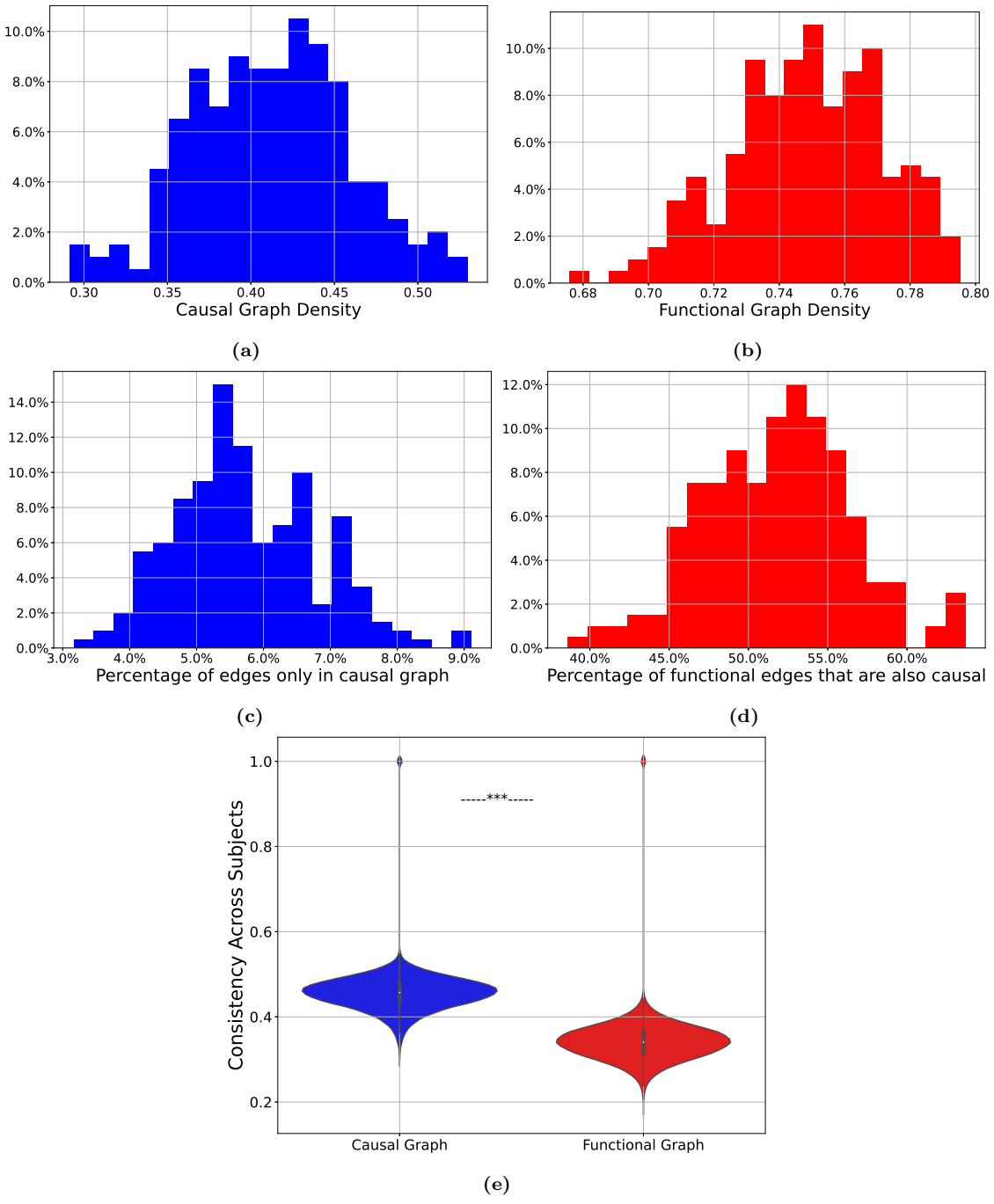


Supplementary Figure 39: Detailed performance curves of Mixed-PCMCI over simulated fMRI from Full Macaque network when using ground-truth parent sets for each conditional independence test. In all graphs, the error bars depict the standard error of the mean.

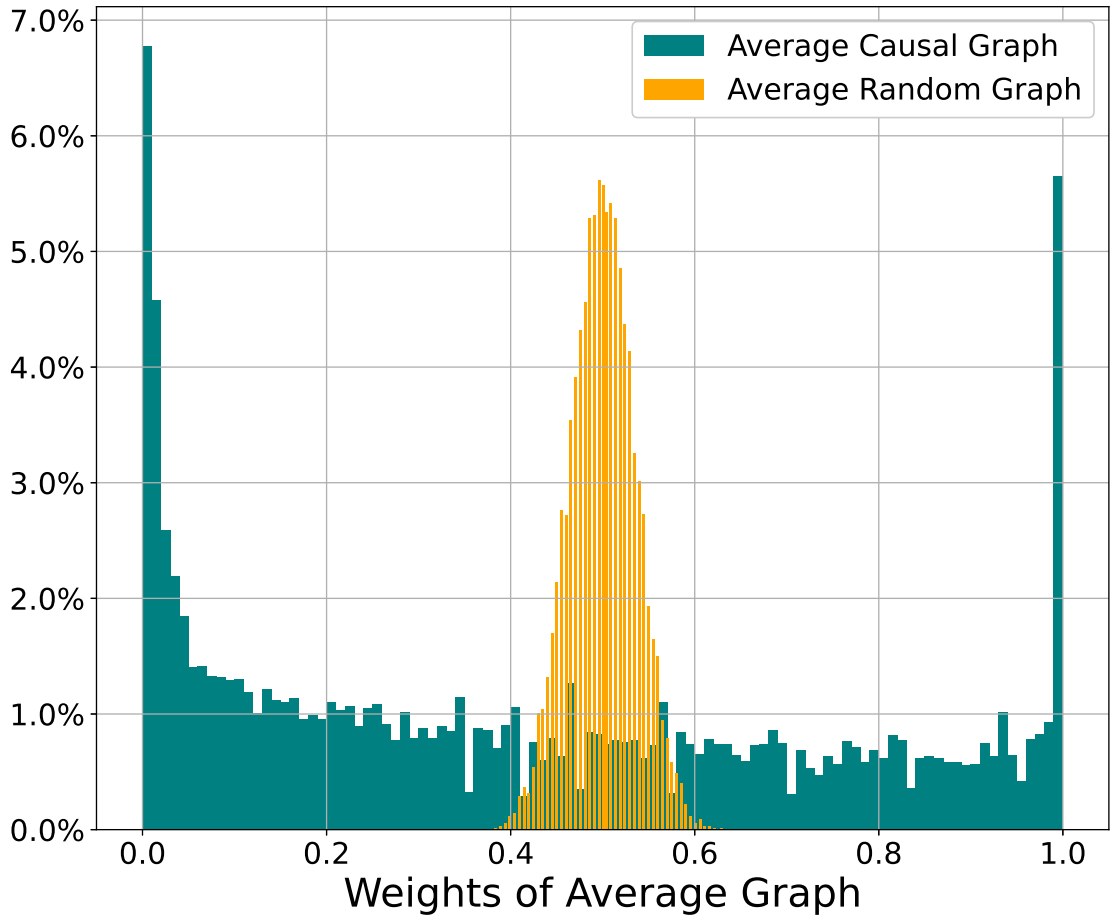


Supplementary Figure 40: Detailed performance curves of CaLLTiF over simulated fMRI from Full Macaque network when using ground-truth parent sets for each conditional independence test. In all graphs, the error bars depict the standard error of the mean.

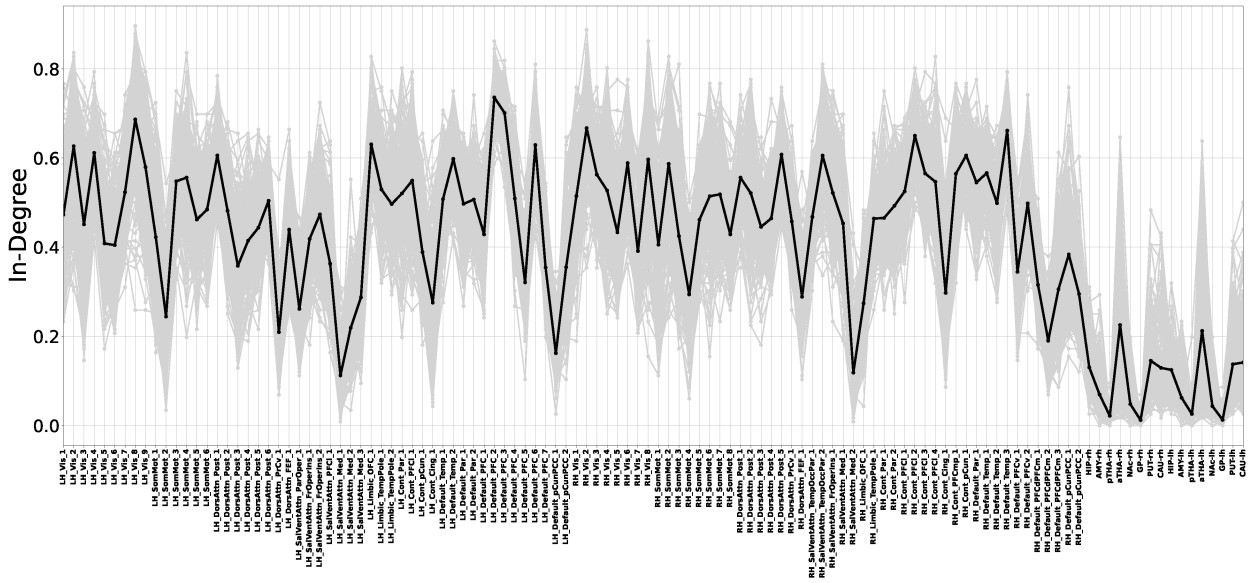
Supplementary Figures for Resting-State Human fMRI from HCP



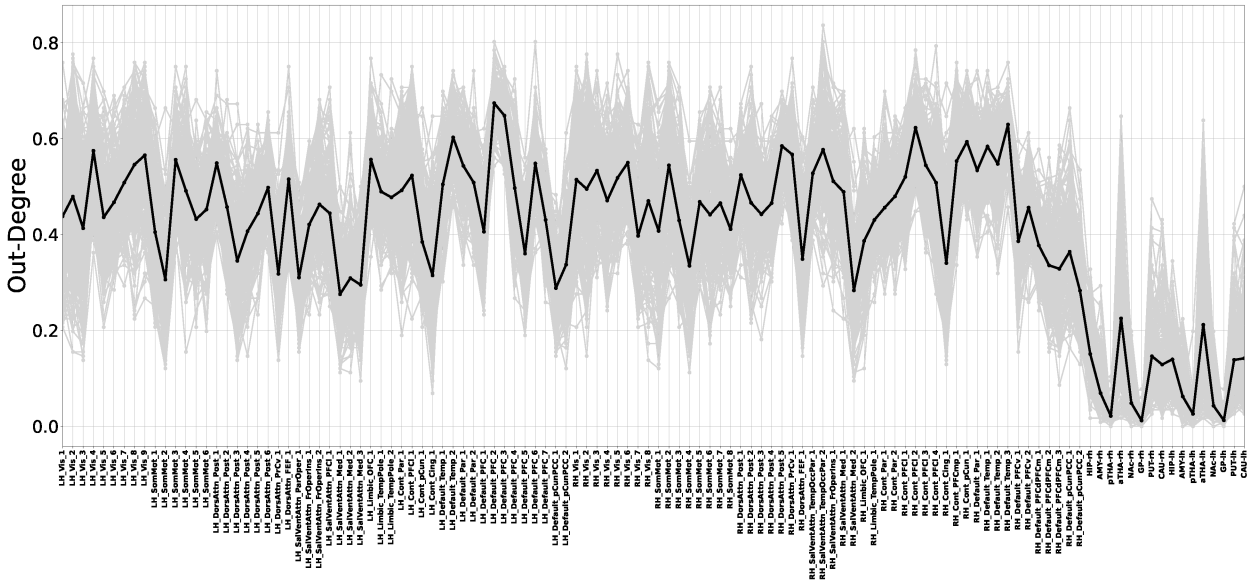
Supplementary Figure 41: Causal graphs are sparse and more consistent than functional connectivity (FC) graphs. (a) Distribution of the density of edges in causal graphs learned by CaLLTiF. (b) Similar to (a) but for FC. (c) Distribution of the percentage of edges that only exist in CaLLTiF but not in FC. (d) Distribution of the percentage of edges that exist in both causal and functional graphs. (e) Distributions of graph consistencies (correlation coefficients) across subjects. $^{***} = p < 0.001$, one-sided Wilcoxon signed-rank test.



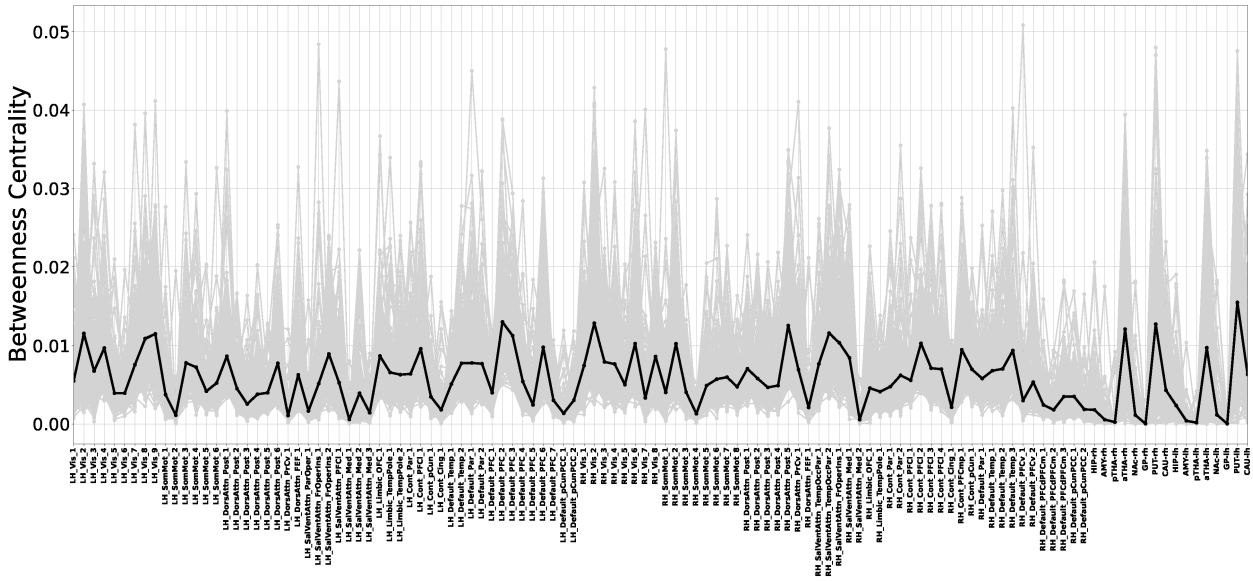
Supplementary Figure 42: Distributions of edge weights of the average causal graph (computed for 200 HCP subjects) and corresponding weights of the average of 200 random binary graphs. The randomized surrogate is computed based on data from 200 subjects, with the weight distribution of the average random graph, which is computed based on data from 200 random binary matrices from the Bernoulli distribution with $p = 0.5$. The average graph based on real data is significantly more bimodal, indicating the presence of evidence in the data on whether each edge exists or not.



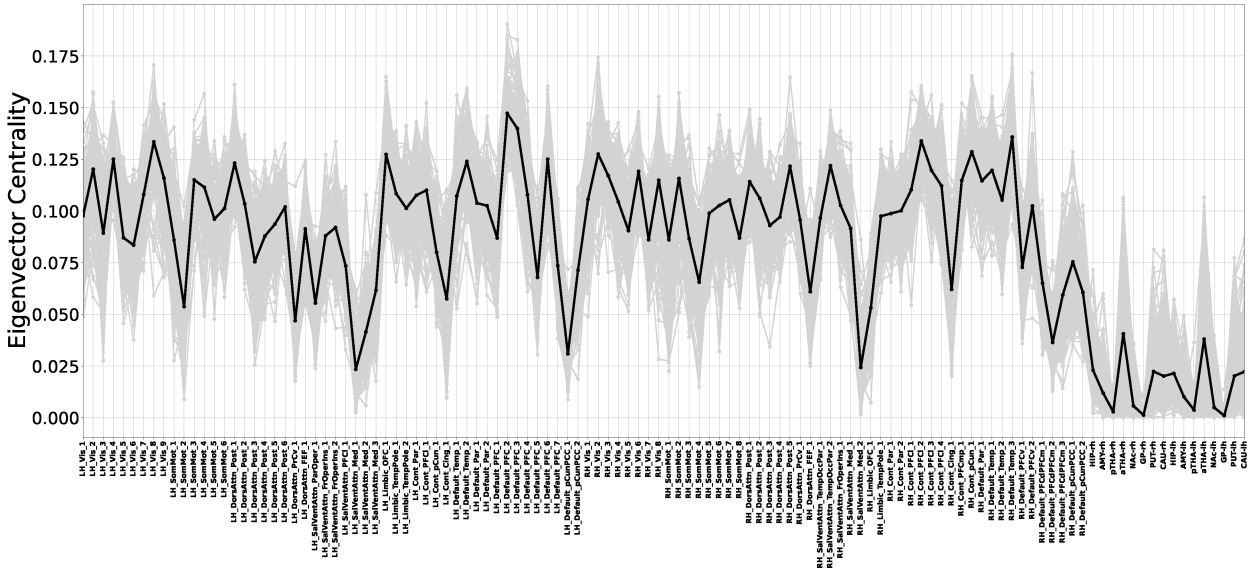
Supplementary Figure 43: In-degree for all nodes (horizontal axis) and all subjects (gray lines) in causal graphs learned by CaLLTiF over human fMRI data from HCP.



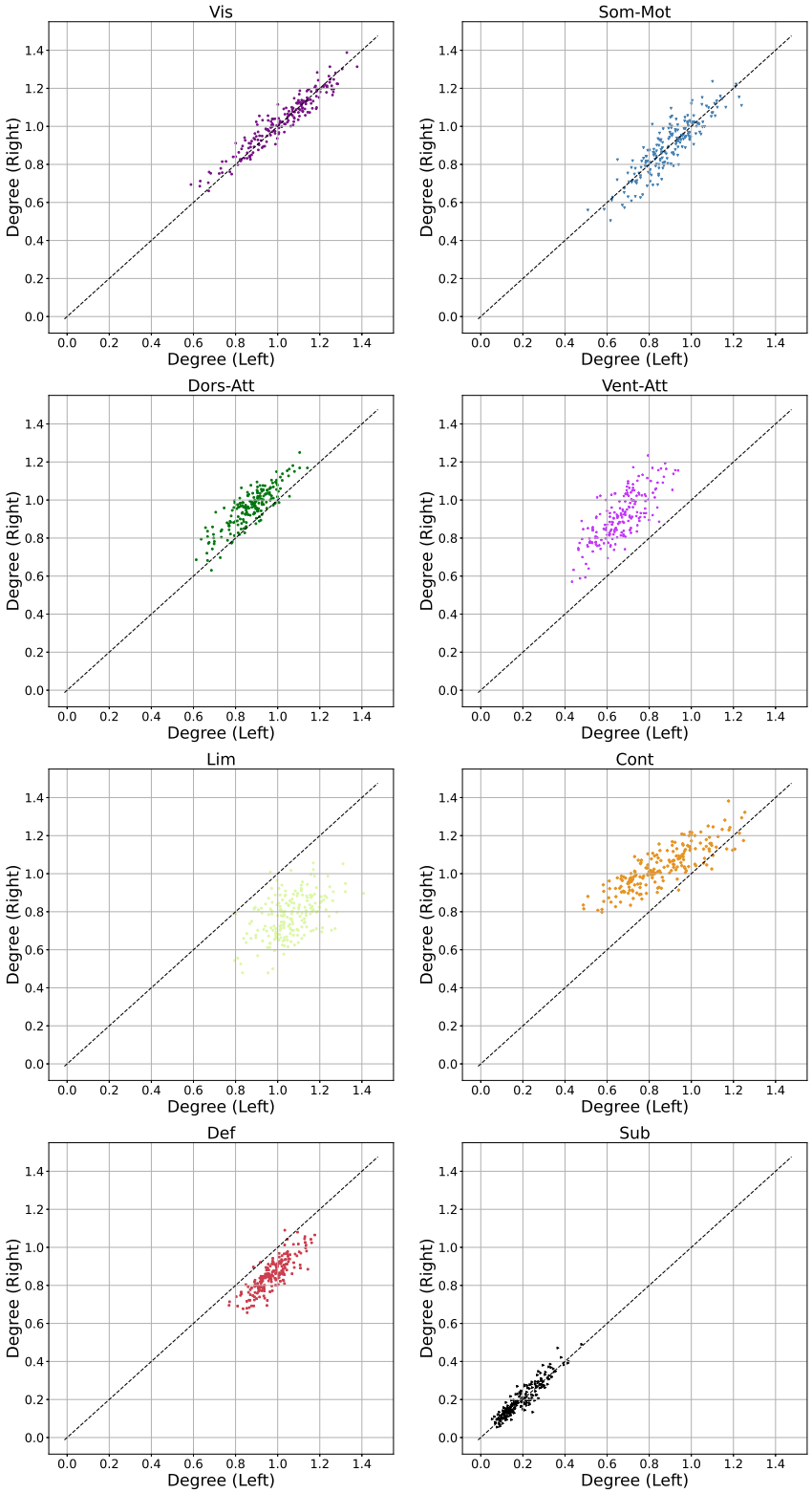
Supplementary Figure 44: Out-degree for all nodes (horizontal axis) and all subjects (gray lines) in causal graphs learned by CaLLTiF over human fMRI data from HCP.



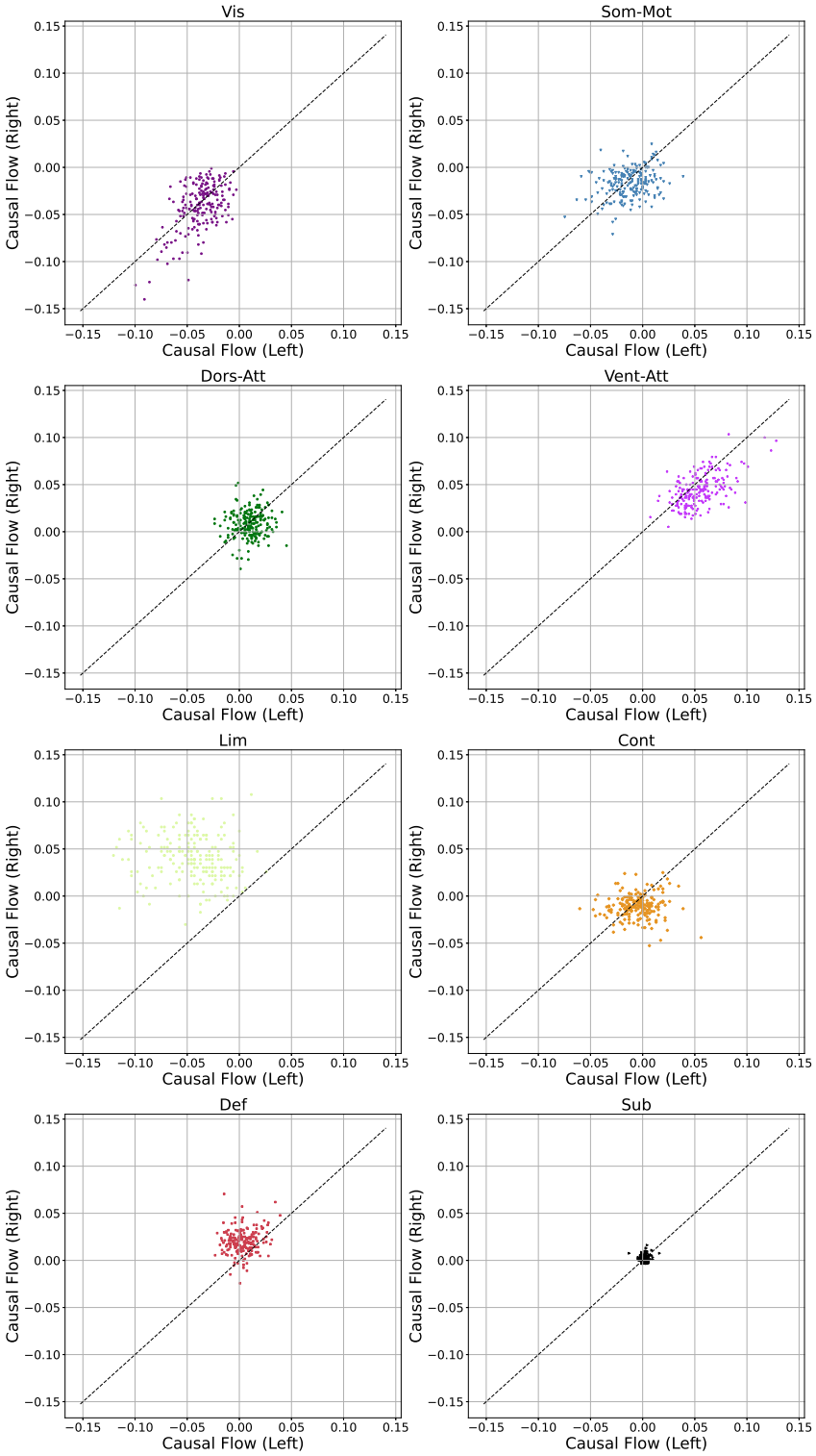
Supplementary Figure 45: Betweenness centrality for all nodes (horizontal axis) and all subjects (gray lines) in causal graphs learned by CaLLTiF over human fMRI data from HCP.



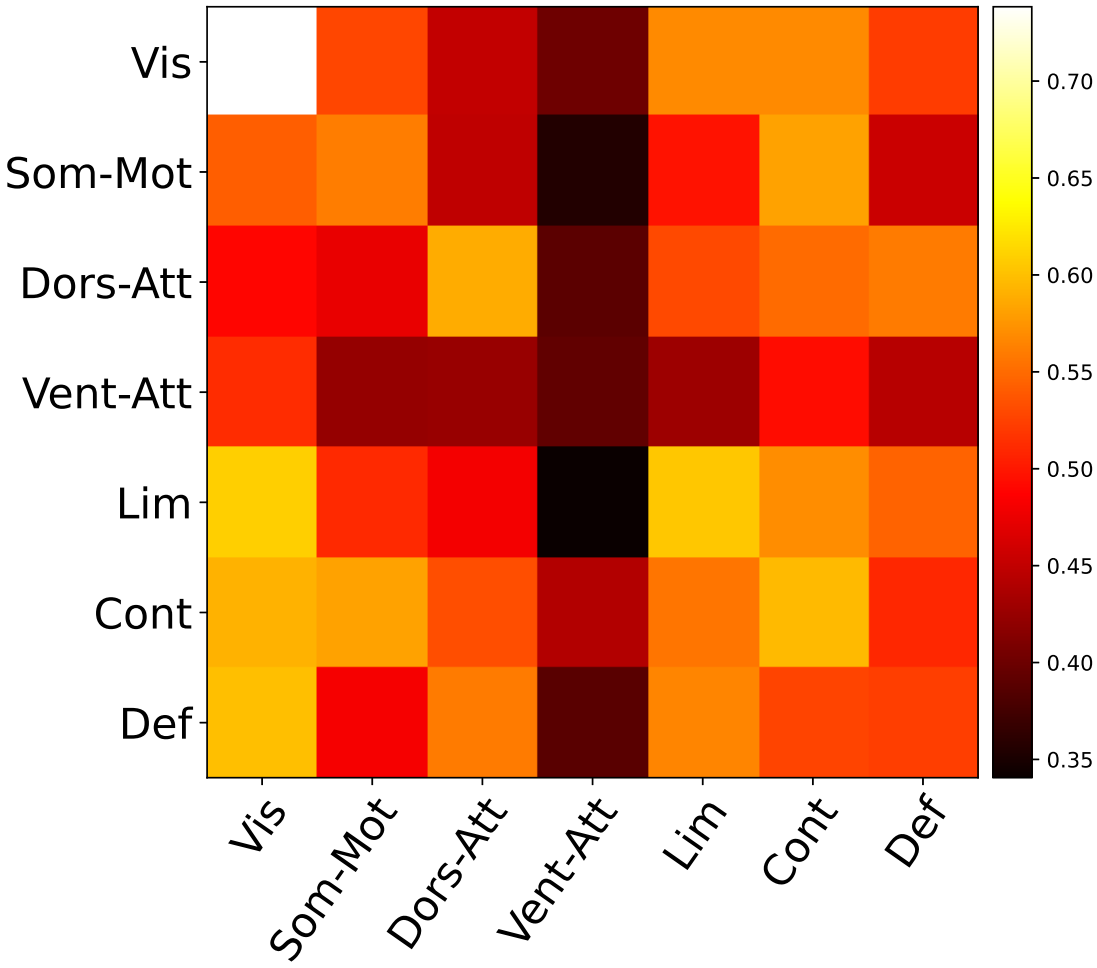
Supplementary Figure 46: Eigenvector centrality for all nodes (horizontal axis) and all subjects (gray lines) in causal graphs learned by CaLLTiF over human fMRI data from HCP.



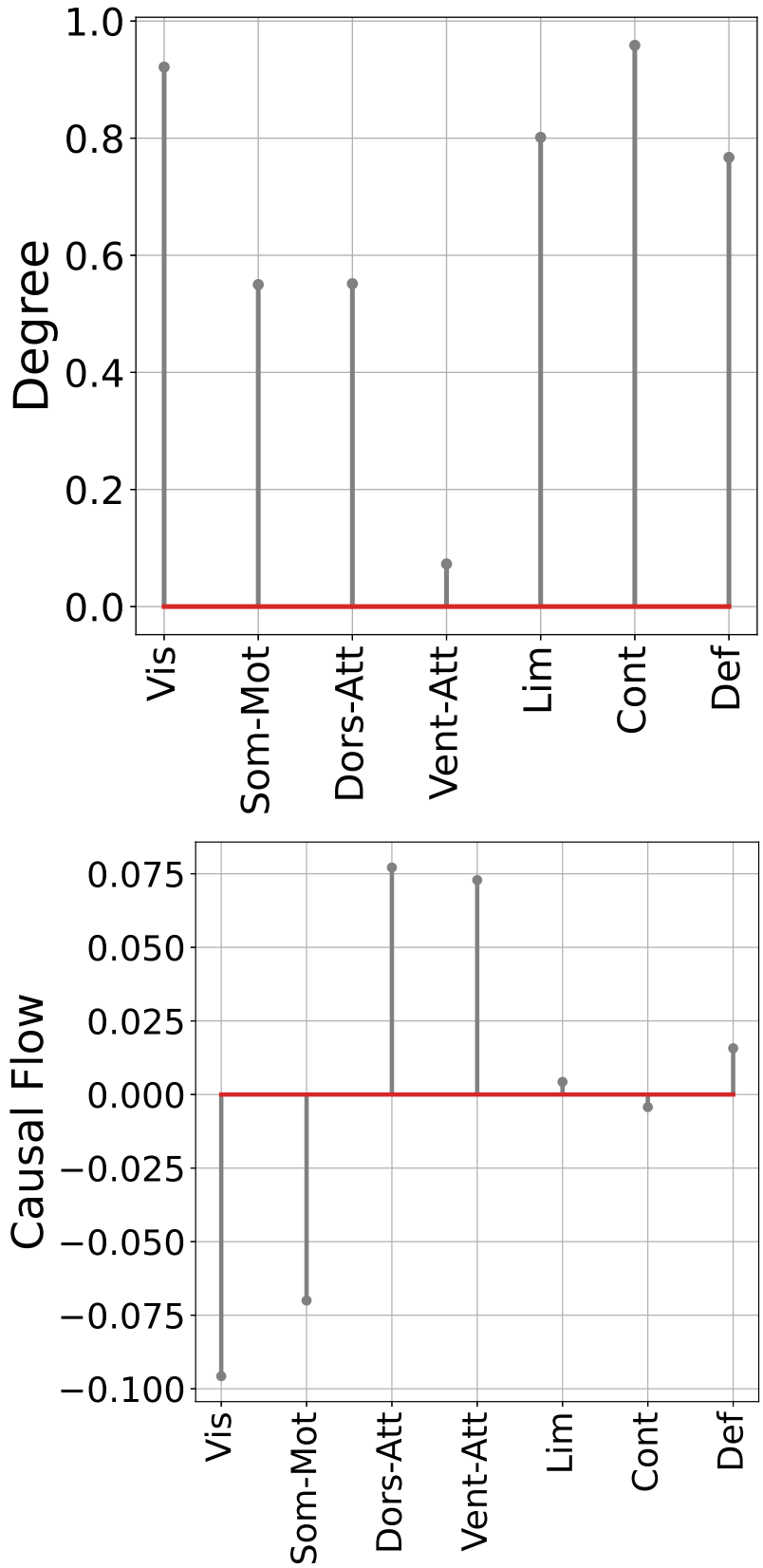
Supplementary Figure 47: Hemispheric (a)symmetry of nodal degrees in causal graphs learned by CaLLTiF over human fMRI data from HCP. This is the same as Figure 8c in the main text except that different subnetworks are shown in distinct panels for better visualization.



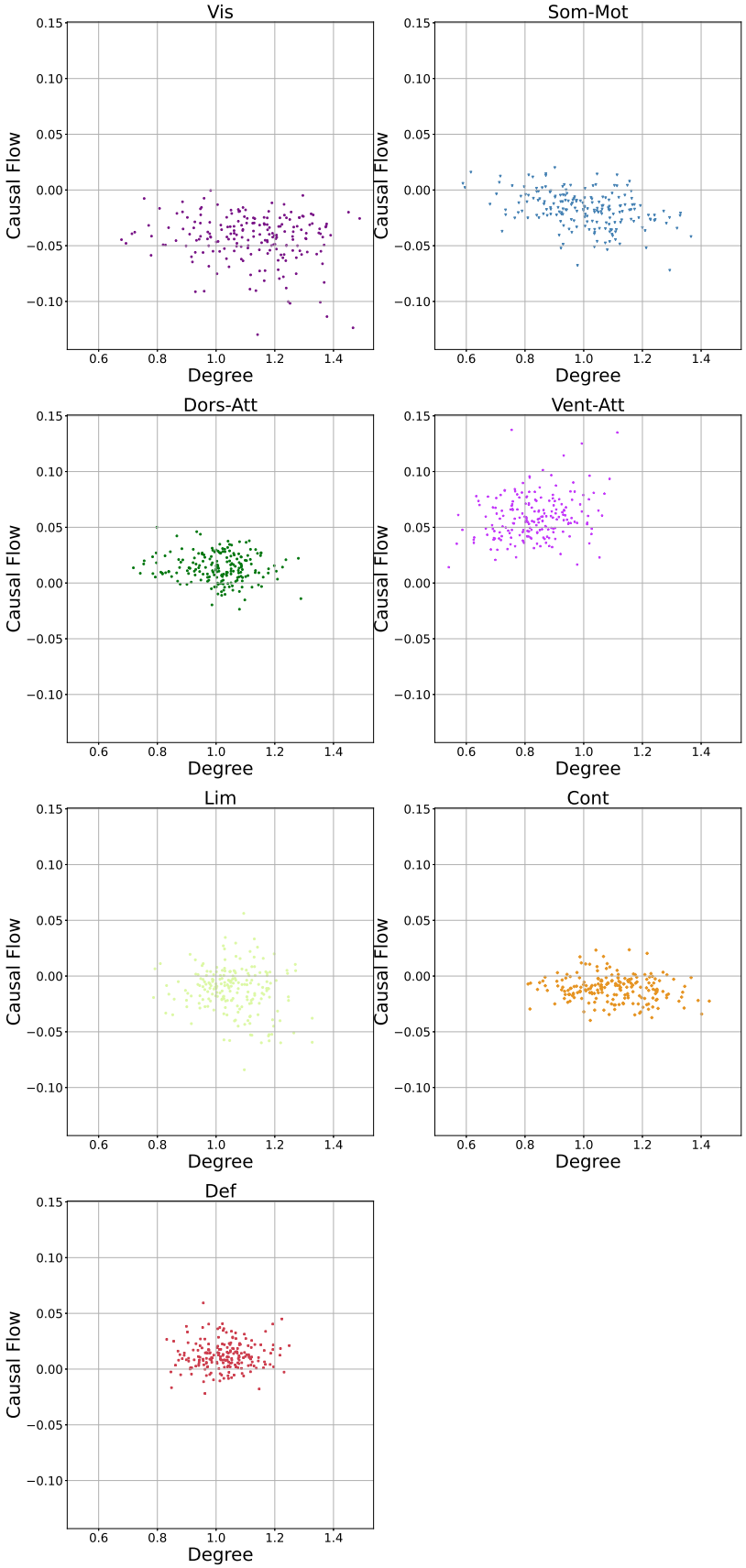
Supplementary Figure 48: Hemispheric (a)symmetry of causal flows in causal graphs learned by CaLLTiF over human fMRI data from HCP. This is the same as Figure 8d in the main text except that different subnetworks are shown in distinct panels for better visualization.



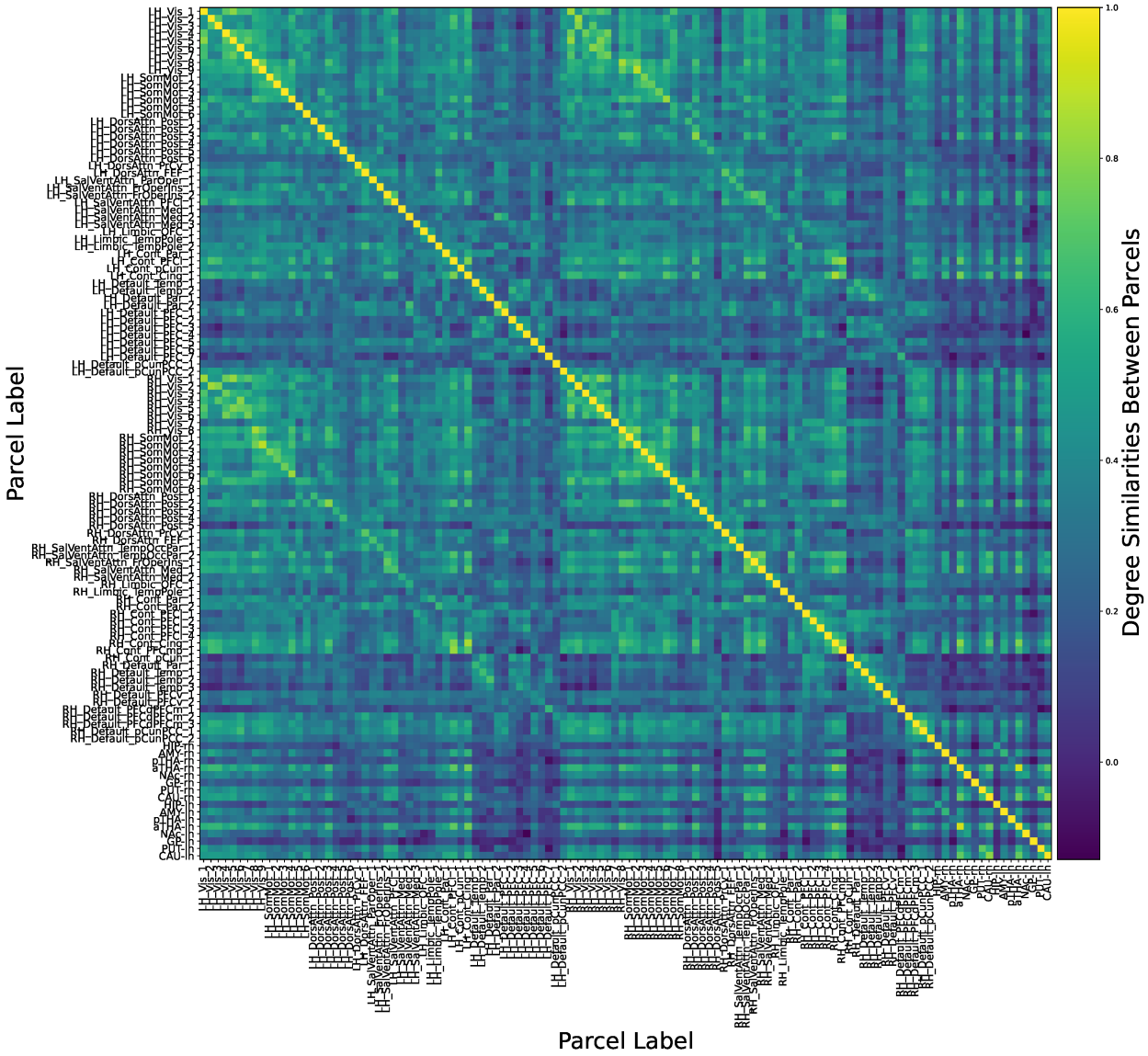
Supplementary Figure 49: **The average subnetwork graph, computed as the mean of subnetwork graphs of all the subjects. In the subnetwork graph of each subject, the weight of an edge from subnetwork i to j is the number of nodes in subnetwork i that connect to nodes in subnetwork j, normalized by the number of all possible edges between these subnetworks.**



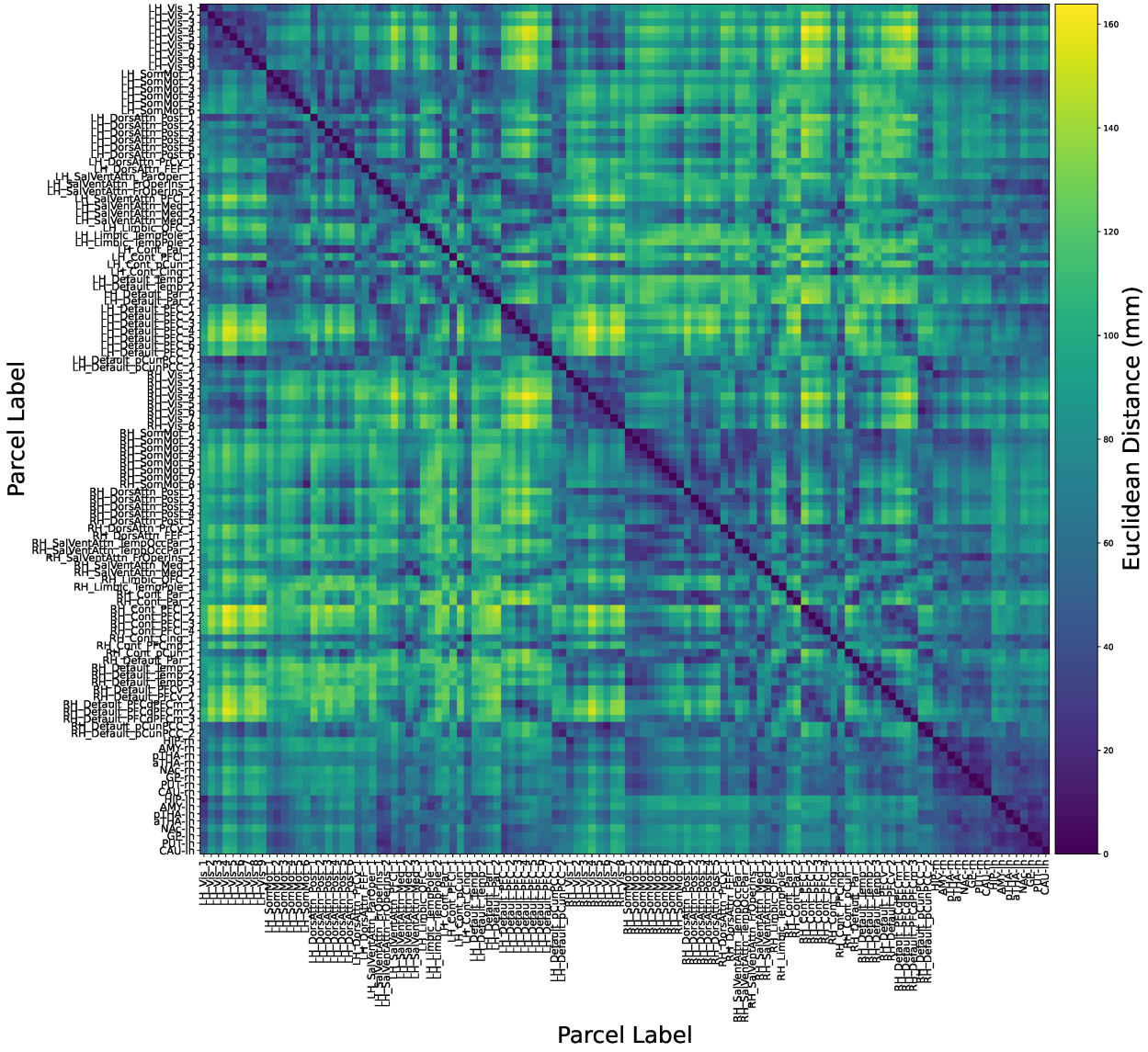
Supplementary Figure 50: Nodal degree and causal flow for each node in the average subnetwork graph in Supplementary Figure 49



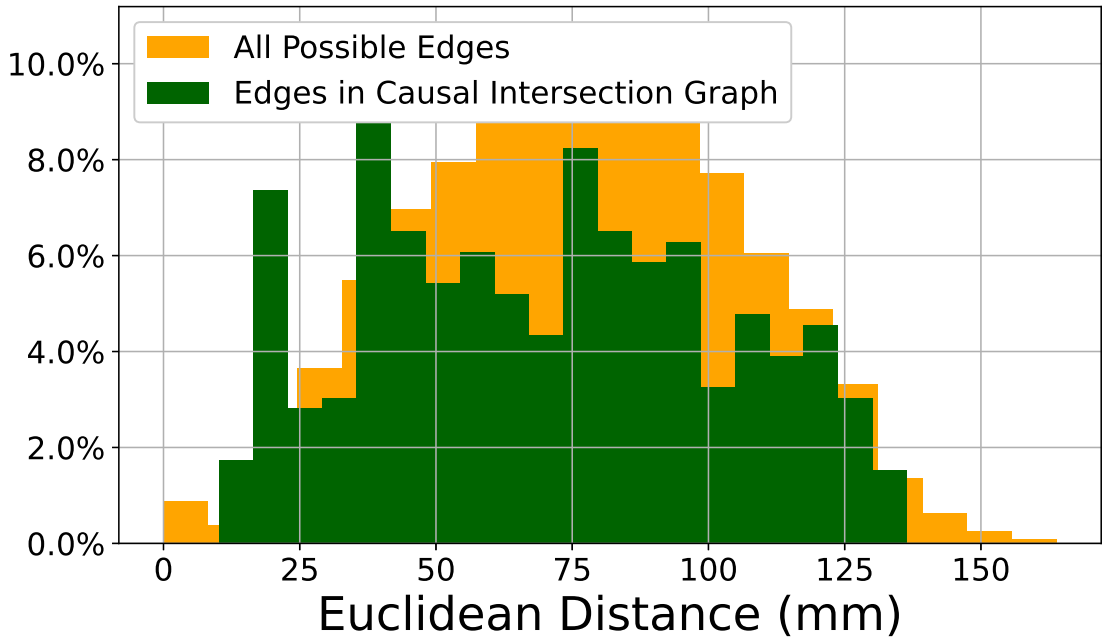
Supplementary Figure 51: Joint distribution of degree and causal flow for (hyper-)nodes of subnetwork graphs for each subject. Each dot in each panel represents one subject and different panels show different functional subnetworks. This is the same as Figure 5b in the main text except that different subnetworks are shown in distinct panels for better visualization.



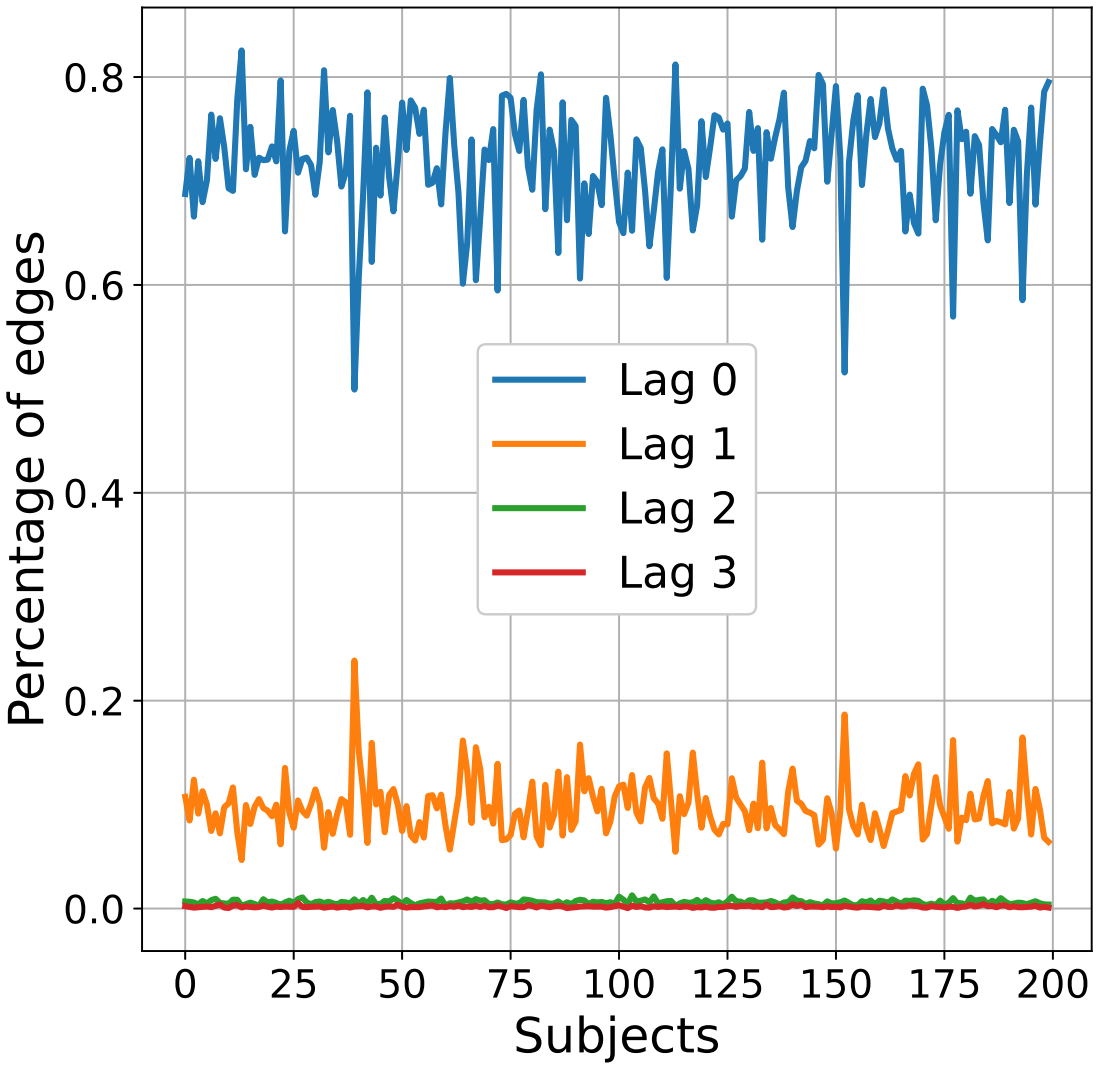
Supplementary Figure 52: Degree similarities between each pair of nodes (parcels) in causal graphs learned by CaLLTiF over human fMRI data from HCP. Each cell shows the correlation coefficient between nodal degrees of the respective two nodes across different subjects.



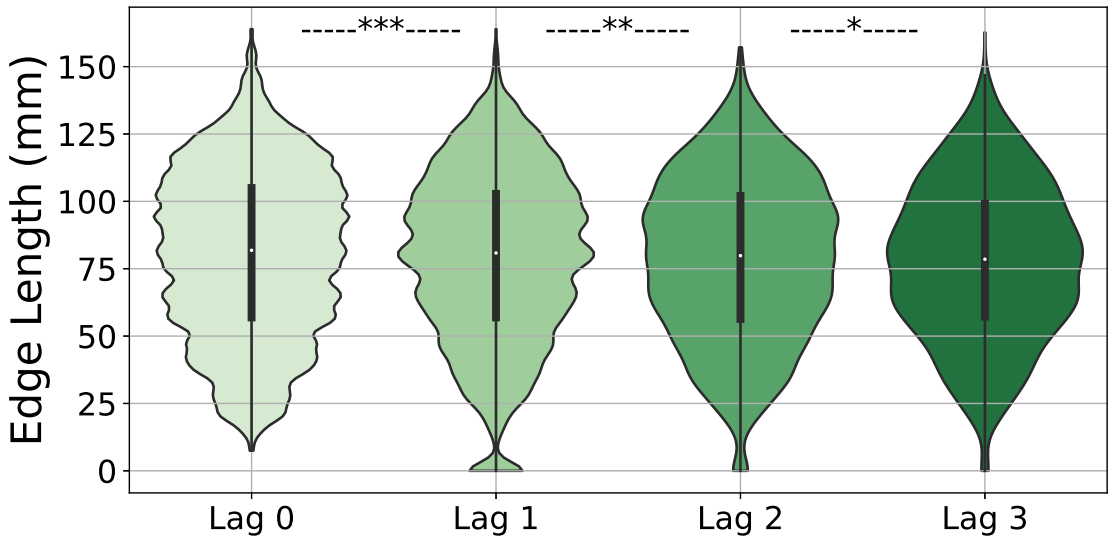
Supplementary Figure 53: Pairwise spatial (Euclidean) distances between all nodes (parcels) of the causal graphs learned by CaLLTiF over human fMRI data from HCP.



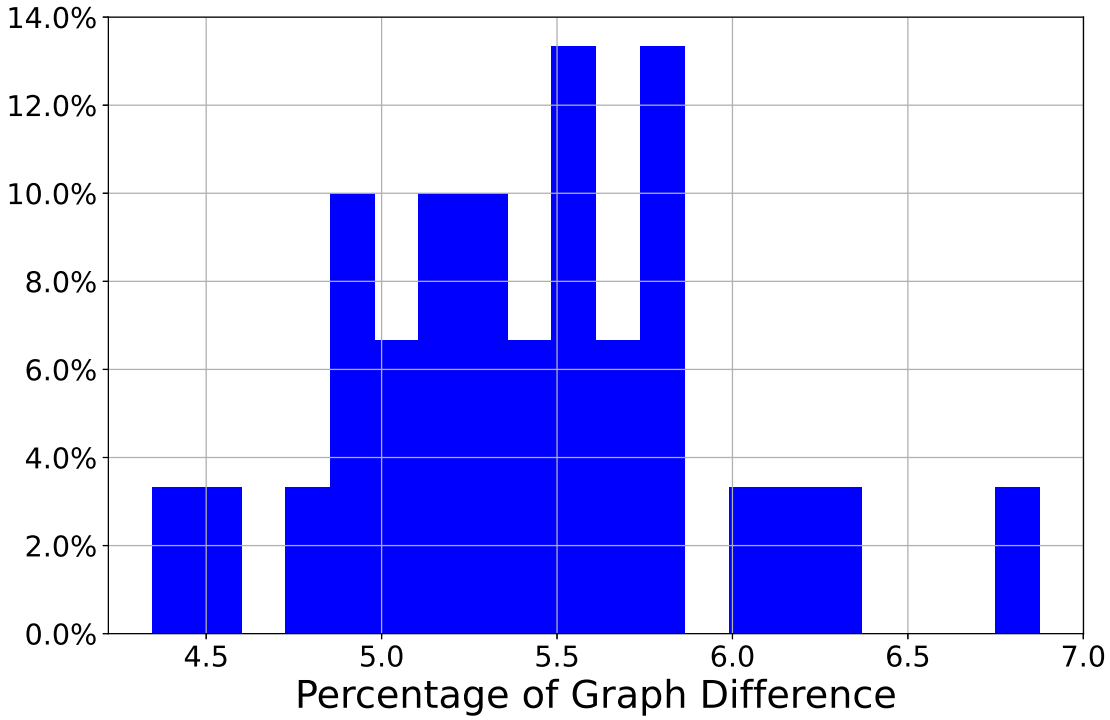
Supplementary Figure 54: Distribution of Euclidean distances for each possible pair of nodes, compared to the distribution of Euclidean distances only for links present in the intersection of all subject-wise causal graphs learned by CaLLTiF over human fMRI data from HCP.



Supplementary Figure 55: Percentage of edges present in each lag of the causal graphs learned by CaLLTiF over human fMRI data from HCP, shown separately for each subject.



Supplementary Figure 56: Comparison between Euclidean edge length of CaLLTiF edges at different lags. Each violin plot shows the distribution of Euclidean edge length (parcel distance between edge endpoints) for edges existing in the subgraph corresponding to each lag. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, one-sided Wilcoxon rank-sum test.



Supplementary Figure 57: Robustness of CaLLTiF on HCP data: distribution of differences between the graphs computed using CaLLTiF with maximum lags of 1 and 3. The difference is computed as the percentage of absolute change in the corresponding binary graphs computed for each subject. The distribution is computed based on the results for 30 subjects.

Table 1: Names of regions in Schaefer 100x7 atlas (cortical parcels) and Melbourne Scale I atlas (subcortical parcels)

Parcel Number	Short Name
1	LH_Vis_1
2	LH_Vis_2
3	LH_Vis_3
4	LH_Vis_4
5	LH_Vis_5
6	LH_Vis_6
7	LH_Vis_7
8	LH_Vis_8
9	LH_Vis_9
10	LH_SomMot_1
11	LH_SomMot_2
12	LH_SomMot_3
13	LH_SomMot_4
14	LH_SomMot_5
15	LH_SomMot_6
16	LH_DorsAttn_Post_1
17	LH_DorsAttn_Post_2
18	LH_DorsAttn_Post_3
19	LH_DorsAttn_Post_4
20	LH_DorsAttn_Post_5
21	LH_DorsAttn_Post_6
22	LH_DorsAttn_PrCv_1
23	LH_DorsAttn_FEF_1
24	LH_SalVentAttn_ParOper_1
25	LH_SalVentAttn_FrOperIns_1
26	LH_SalVentAttn_FrOperIns_2
27	LH_SalVentAttn_PFCl_1
28	LH_SalVentAttn_Med_1
29	LH_SalVentAttn_Med_2
30	LH_SalVentAttn_Med_3
31	LH_Limbic_OFC_1
32	LH_Limbic_TempPole_1
33	LH_Limbic_TempPole_2
34	LH_Cont_Par_1
35	LH_Cont_PFCl_1
36	LH_Cont_pCun_1
37	LH_Cont_Cing_1
38	LH_Default_Temp_1
39	LH_Default_Temp_2
40	LH_Default_Par_1
41	LH_Default_Par_2
42	LH_Default_PFC_1
43	LH_Default_PFC_2

Continued on next page

Table 1 – *Continued from previous page*

Parcel Number	Short Name
44	LH_Default_PFC_3
45	LH_Default_PFC_4
46	LH_Default_PFC_5
47	LH_Default_PFC_6
48	LH_Default_PFC_7
49	LH_Default_pCunPCC_1
50	LH_Default_pCunPCC_2
51	RH_Vis_1
52	RH_Vis_2
53	RH_Vis_3
54	RH_Vis_4
55	RH_Vis_5
56	RH_Vis_6
57	RH_Vis_7
58	RH_Vis_8
59	RH_SomMot_1
60	RH_SomMot_2
61	RH_SomMot_3
62	RH_SomMot_4
63	RH_SomMot_5
64	RH_SomMot_6
65	RH_SomMot_7
66	RH_SomMot_8
67	RH_DorsAttn_Post_1
68	RH_DorsAttn_Post_2
69	RH_DorsAttn_Post_3
70	RH_DorsAttn_Post_4
71	RH_DorsAttn_Post_5
72	RH_DorsAttn_PrCv_1
73	RH_DorsAttn_FEF_1
74	RH_SalVentAttn_TempOccPar_1
75	RH_SalVentAttn_TempOccPar_2
76	RH_SalVentAttn_FrOperIns_1
77	RH_SalVentAttn_Med_1
78	RH_SalVentAttn_Med_2
79	RH_Limbic_OFC_1
80	RH_Limbic_TempPole_1
81	RH_Cont_Par_1
82	RH_Cont_Par_2
83	RH_Cont_PFCl_1
84	RH_Cont_PFCl_2
85	RH_Cont_PFCl_3
86	RH_Cont_PFCl_4
87	RH_Cont_Cing_1
88	RH_Cont_PFCmp_1
89	RH_Cont_pCun_1

Continued on next page

Table 1 – *Continued from previous page*

Parcel Number	**Short Name**
90	RH_Default_Par_1
91	RH_Default_Temp_1
92	RH_Default_Temp_2
93	RH_Default_Temp_3
94	RH_Default_PFCv_1
95	RH_Default_PFCv_2
96	RH_Default_PFCdPFCm_1
97	RH_Default_PFCdPFCm_2
98	RH_Default_PFCdPFCm_3
99	RH_Default_pCunPCC_1
100	RH_Default_pCunPCC_2
101	HIP-rh
102	AMY-rh
103	pTHA-rh
104	aTHA-rh
105	NAc-rh
106	GP-rh
107	PUT-rh
108	CAU-rh
109	HIP-lh
110	AMY-lh
111	pTHA-lh
112	aTHA-lh
113	NAc-lh
114	GP-lh
115	PUT-lh
116	CAU-lh