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 decades of research, discovering causal relationships from purely observational neuroimaging data such as fMRI remains a challenge. Popular algorithms such as Multivariate Granger Causality (MVGC) and Dynamic Causal Modeling (DCM) fall short in handling complex aspects of data such as contemporaneous effects and latent common causes. Decades of research on causal structure learning have developed alternative algorithms that address these limitations, but they often scale poorly with the number of variables and rely heavily on the lack of cycles in the underlying graph. Further, how existing algorithms compare in terms of accuracy and scalability when applied to fMRI has remained unknown. In this work, we first provide a detailed analysis of existing methods, finding Runge et al's PCMCI algorithm [1] and MVGC most accurate over simulated fMRI. However, neither algorithm is able to detect directed contemporaneous effects, a capability that is particularly important for fMRI. To address this gap, we propose the Causal discovery for Large-scale Low-resolution Time-series with Feedback (CaLLTiF) algorithm. This algorithm is based on PCMCI but improves upon it in terms of computational efficiency, cyclic contemporaneous connections, and statistical type I error control. CaLLTiF achieves significantly higher accuracy compared to existing algorithms in simulated fMRI. When applied to resting-state fMRI from human subjects ($n = 200$, Human Connectome Project), causal connectomes learned by CaLLTiF show higher sparsity and consistency across subjects than functional connectivity and align with known resting-state dynamics. These graphs also capture Euclidean distance-dependence in causal interactions, while demonstrating statistically significant laterality and gender differences. Overall, this work provides a clear picture of the critical gap between the capabilities of existing algorithms and the needs of causal discovery from whole-brain fMRI, and proposes a new algorithm with higher scalability and accuracy to address this gap. This study further paves the way and defines a standard for future investigations into causal discovery from neuroimaging data.

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

Introduction

A major step in the global drive for understanding the brain [2–6] is to move beyond correlations and understand the causal relationships among internal and external factors – a process often referred to as *causal discovery* [7,8]. 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) [9,10]. 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 [11–13], but also because of the great extent to which

the presence of unobserved variables can complicate delineating causal adjacencies and orientations [14–17]. Nevertheless, many characteristics of fMRI also make causal discovery challenging, including its large dimensionality, low temporal resolution, and indirect reflection of underlying neural processes [18]. 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). GC has a long history in neuroscience [19, 20], 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 [21], which is about one order of magnitude slower than the time that it takes for neural signals to travel across the brain [22–24]. 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 [25, Appendix A]). Such fast sub-TR interactions demonstrate themselves as causal effects that appear to be "contemporaneous" (i.e., instantaneous) and can even be cyclic, making causal discovery significantly more challenging. Similar to GC, Dynamic Causal Modeling (DCM) has also been widely used with fMRI data [26–29] 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 [30–32].

Discovering causal relationships without reliance on time has been the subject of extensive research in the causal inference literature [7, 33–37]. A wide range of algorithmic solutions have been proposed [7, 18, 36, 38–45], which are often classified based on their methodology into constraint-based [33, 34, 46], noise-based [43, 47], and score-based [48, 49]. 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 analysis based on their theoretical characterizations and numerical performance on multiple simulated fMRI benchmarks. This discussion demonstrates a large gap between what causal discovery for fMRI needs and what existing algorithms can achieve. To address this gap, we propose the design of a new algorithm called *Causal discovery for Large-scale Low-resolution Time-series with Feedback (CaLLTiF)*. We demonstrate the higher accuracy of CaLLTiF against state-of-the-art algorithms in simulated fMRI and provide a case study of its performance on resting-state fMRI data from the Human Connectome Project (HCP) [50].

Results

A Taxonomy of Causal Discovery from 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) [20, 51] does not does not satisfy criteria (2), but we still included it in our subsequent analyses due to its popularity in neuroscience [52–55]. On the other hand, we excluded LiNG [56] from further analysis since it is considered by its proposers as generally inferior to FASK [57].

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 [57]. 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). Figures 1b and 1c illustrate the distribution of F1 scores for all algorithms, combined across nine simple networks. The results show that the PCMCI algorithm achieved the highest median F1 score, both over the directed graphs (all z > 4.98 and $p < 10^{-6}$, Wilcoxon signed-rank test, computed between PCMCI and each method separately) and undirected skeletons (all z > 11.32 and $p < 10^{-29}$,

Wilcoxon signed-rank test, computed between PCMCI and each method separately). The PCMCI algorithm also has the smallest computational complexity on simple networks, as seen from Figure 1d. Furthermore, our results indicate that PCMCI$^+$, FASK, and DGlearn (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 4, 6, and 9).

Next, we compared the remaining four algorithms (PCMCI, MVGC, DYNOTEARS, and VARLiNGAM) on a larger-scale, more realistic simulated benchmark. The graph, called 'Small-Degree Macaque' network, consists of a complete macaque anatomical connectome with 28 nodes and 52 directed edges [57] but the generative model used to simulate fMRI data from this graph remains the same (see Methods for details). The distribution of F1 scores are shown in Figures 2a and 2b. PCMCI and MVGC achieved very similar success in learning both the full graph and its undirected skeleton, while significantly outperforming DYNOTEARS and VARLiNGAM. As far as execution time is concerned, however, MVGC showed 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 also see Figure 4d).

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. As seen from Figure 3a, 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. 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 accuracy of PCMCI monotonically increases with PC Alpha, reaching its maximum at PC Alpha = 1 (Figures 3b and 3c). 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 [1]) 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, and 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

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 [1], PCMCI$^+$ [58]	VARLiNGAM [59]	DYNOTEARS [60]
Cross-sectional with cycles	FASK [57]	FASK [57], LiNG [61]	DGlearn [62]

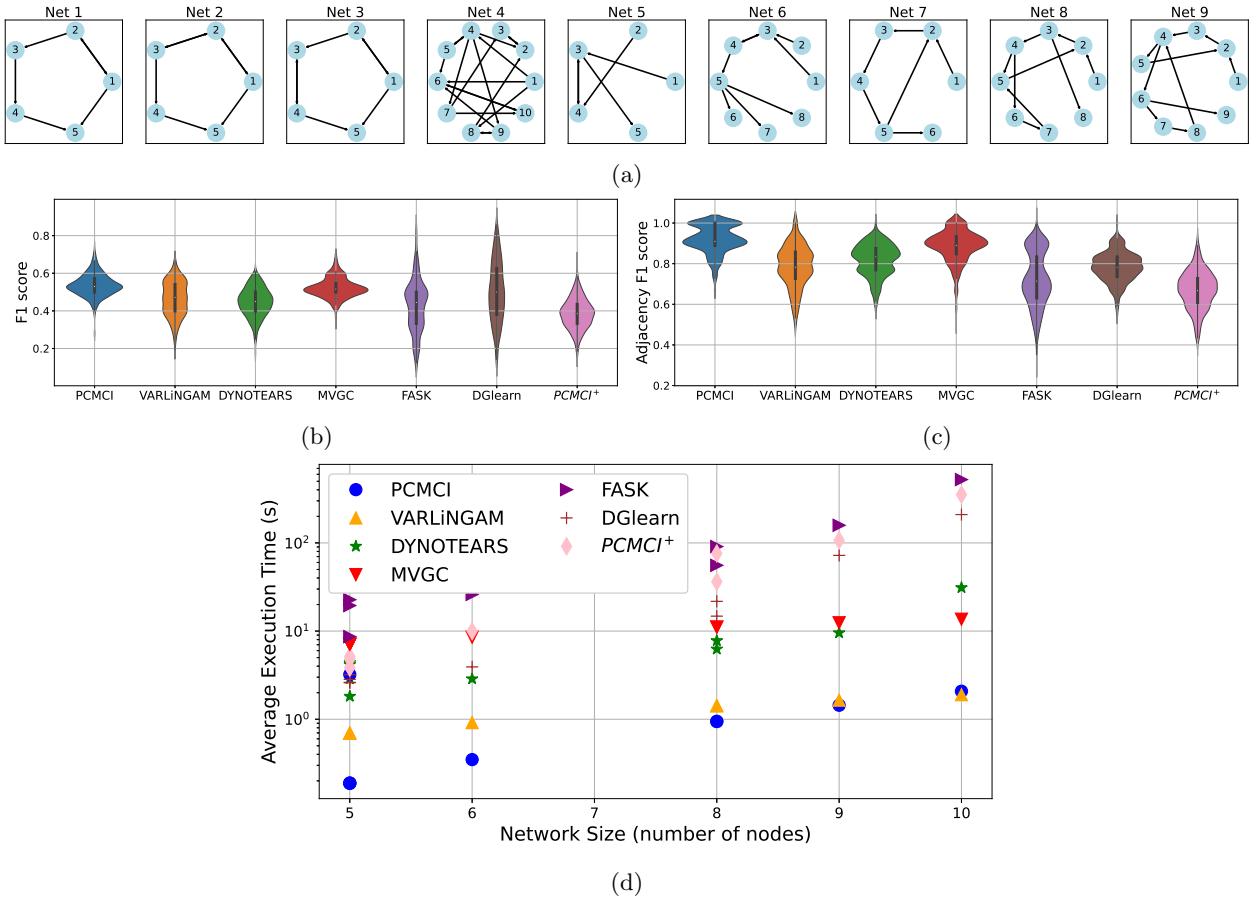


Figure 1: **Results of comparing several state-of-the-art causal discovery algorithms over a benchmark of simulated fMRI [57] generated from simple networks with 5-10 nodes each.** (a) Ground truth graphs of the simple networks in the benchmark. Despite all being small-scale, the graphs vary widely in their density, number of cycles, etc. (b,c) F1 scores 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 F1 score samples. The PCMCI method achieves the highest median F1 score, both directed and undirected. (d) 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.

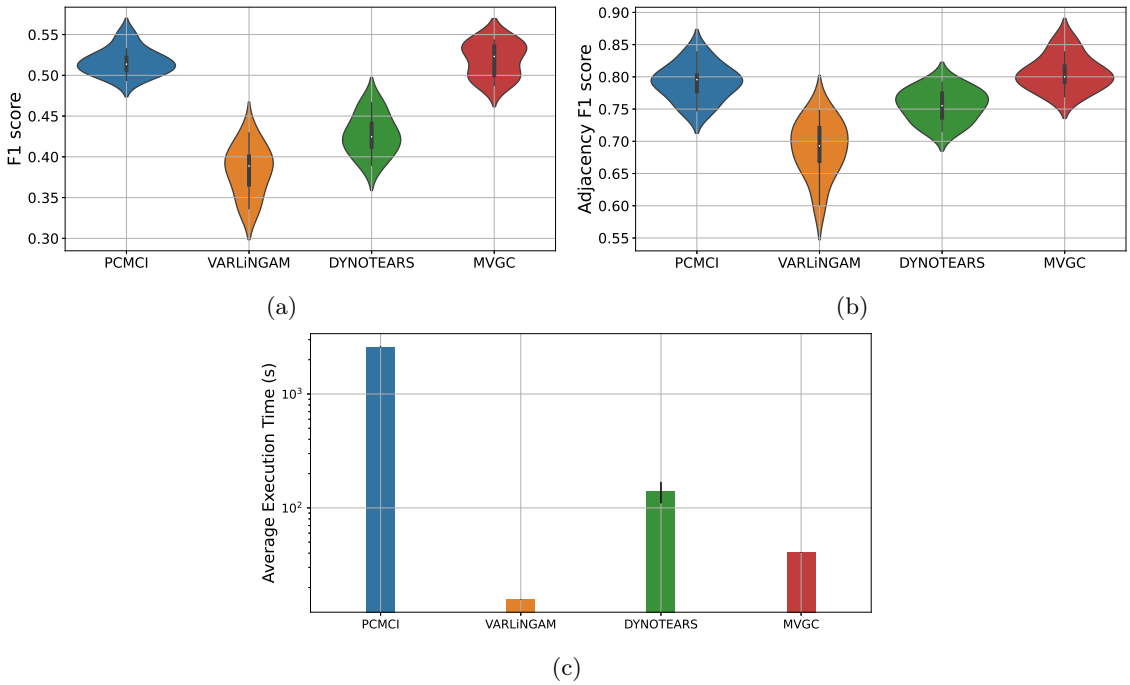


Figure 2: **Comparing the scalable subset of algorithms from Figure 1 over simulated fMRI data from the Small Degree Macaque benchmark [57].** (a) F1 scores of identifying the full (directed) graph. Each distribution consists of 10 F1 scores calculated based on 10 repetitions of simulated data from the same underlying graph. (b) Corresponding F1 scores of identifying the undirected skeleton of the Small Degree Macaque network. (c) Mean execution times for each method (error bars show one standard deviation).

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 [63–65]. 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 4 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 (cf. Methods). CaLLTiF has a significantly higher F1 score (Figure 4a, $p < 10^{-6}$, Wilcoxon signed-rank test) and adjacency F1 score (Figure 4b, $p < 10^{-6}$, Wilcoxon signed-rank test) compared to PCMCI (where only lagged edges are used to construct the summary graph). To further elucidate the benefits of including contemporaneous effects, 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 (see Methods). As expected, Mixed-PCMCI benefits from contemporaneous effects as much as CaLLTiF in adjacency F1 score but not in full F1 score, further motivating the inclusion of directed contemporaneous connections as done in CaLLTiF. Finally, CaLLTiF also achieved significantly higher F1 score (Figure 4c, all $p < 10^{-6}$, Wilcoxon signed-rank test, computed between CaLLTiF and each method separately) and adjacency F1 score (Figure 4d, all $p < 10^{-6}$, Wilcoxon signed-rank test, computed between CaLLTiF and each method separately) than VARLiNGAM, DYNOTEARS, and MVGC in the Full Macaque dataset.

Finally, the third aspect in which CaLLTiF departs from PCMCI is the choice of the 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. Figure 3). 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

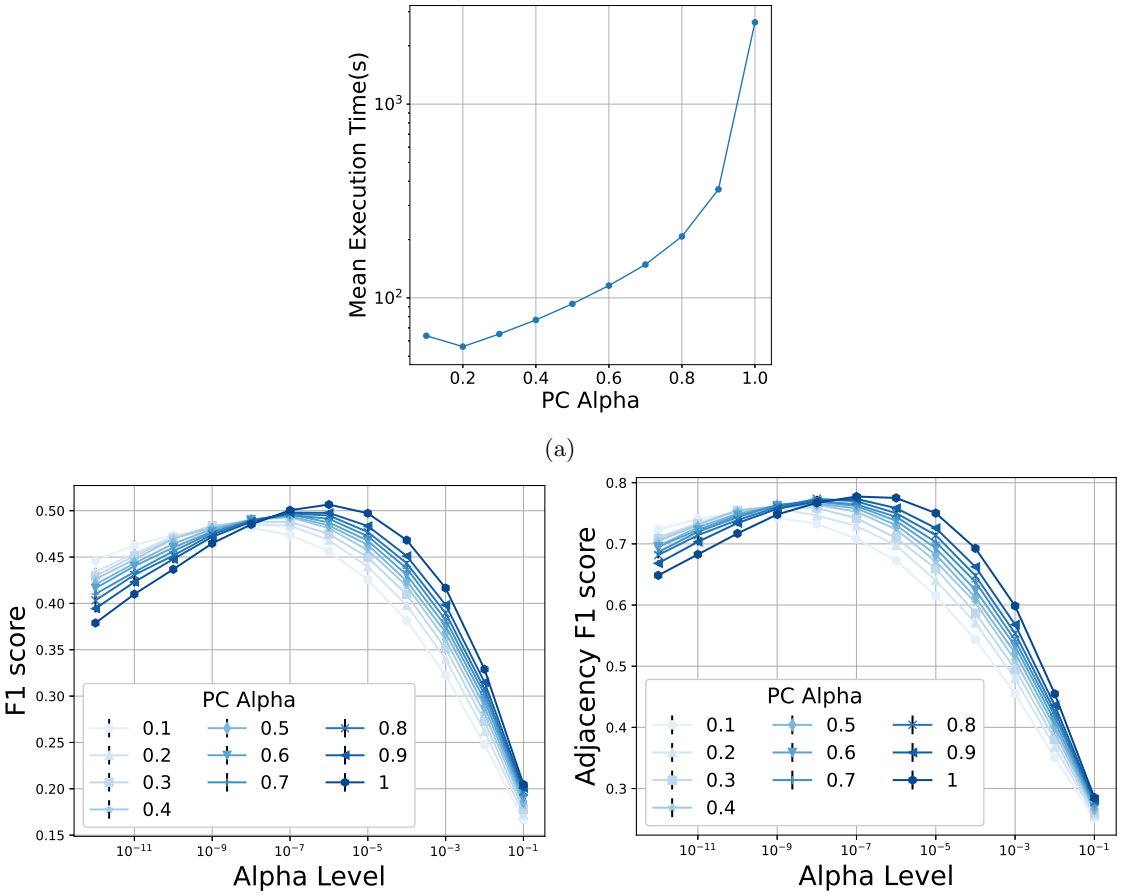


Figure 3: **The effect of the 'PC Alpha' hyperparameter of PCMCI on its performance over simulated Small-Degree Macaque fMRI.** (a) Mean Execution time for different values of PC Alpha, showing a super-exponential growth. (b,c) F1 scores for varying values of PC Alpha and Alpha Level. Alpha Level denotes 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.

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 $X_i(t-\tau) \rightarrow X_j(t)$. 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 to a summary graph by taking an OR operation for each edge across all lags (cf. Methods for details).

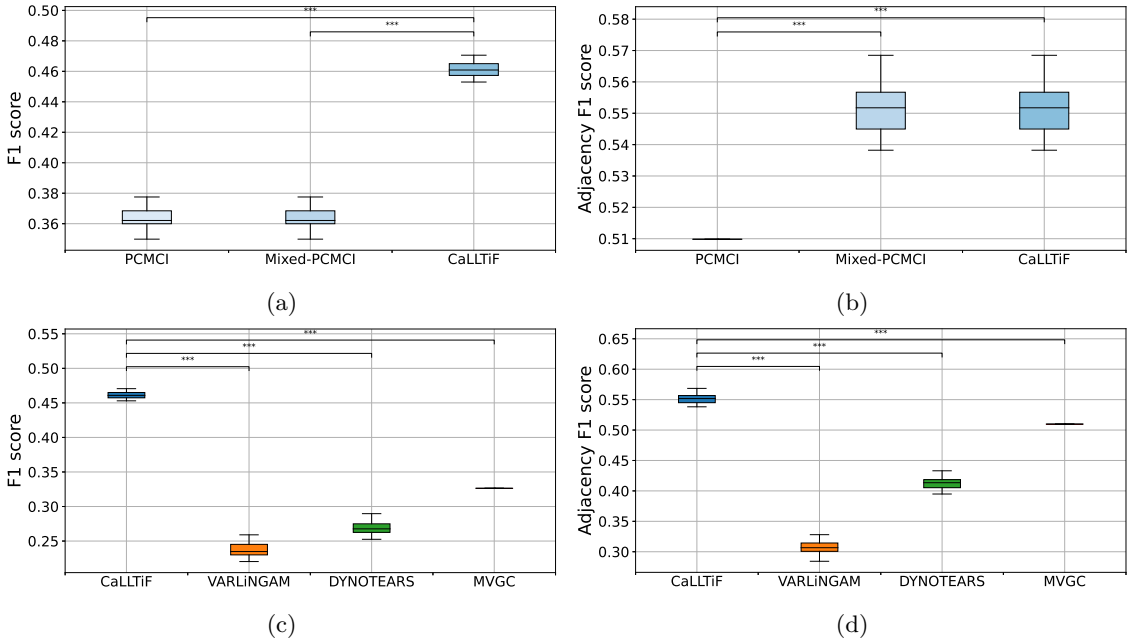


Figure 4: **Comparisons between the proposed algorithm (CaLLTiF) and state-of-the-art alternatives over simulated fMRI from the Full Macaque connectome.** (a,b) Distributions of F1 scores for PCMCI (ignoring the contemporaneous ⧟ connections), Mixed-PCMCI (using the contemporaneous ⧟ connections only for adjacency), and CaLLTiF. (c,d) Distributions of F1 scores for CaLLTiF and state-of-the-art alternatives. 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 Wilcoxon signed-rank test.

Case Study: Whole-Brain Causal Discovery from Resting-State Human fMRI

In this section, we assess the performance of the proposed causal discovery method, CaLLTiF, on a real fMRI dataset. The sample consisted of 200 subjects selected as those with the least head motion from the HCP resting-state fMRI dataset (see Methods). Resting-state scans were parcellated into 100 cortical and 16 subcortical regions (see Methods), and CaLLTiF was performed on all four resting-state scans for each subject, resulting in one causal graph per individual. As noted earlier and explained in Methods, the conservative correction for multiple comparisons across lags in CaLLTiF results in at least 99% statistical confidence in each edge in the resulting graphs. Overall, we found the obtained causal graphs to be 30-55% dense across all subjects (Figure 10a) and 40-60% dense among cortical nodes. These are consistent with about 66% cortical density found using tract-tracing results in non-human primates [65], though slightly lower potentially due to conservative statistical thresholds in CaLLTiF. In the following, we highlight some of the most notable findings from the obtained causal connectomes. Due to a lack of ground-truth connectivity, we analyze the graphs at a population level and examine similarities and differences in graph statistics across individuals.

Learned causal graphs are highly consistent across subjects. Despite individual differences, a remarkably common causal connectome emerged across subjects. Figure 5b shows the average causal graph among the subjects and Figures 5c,5a show 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. We observe significant commonalities among subjects in both the average and the intersection graphs. One can determine the consistency of causal connections among subjects by comparing the weight distribution of the average causal graph, which 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 ran including 461 non-self-loops. As

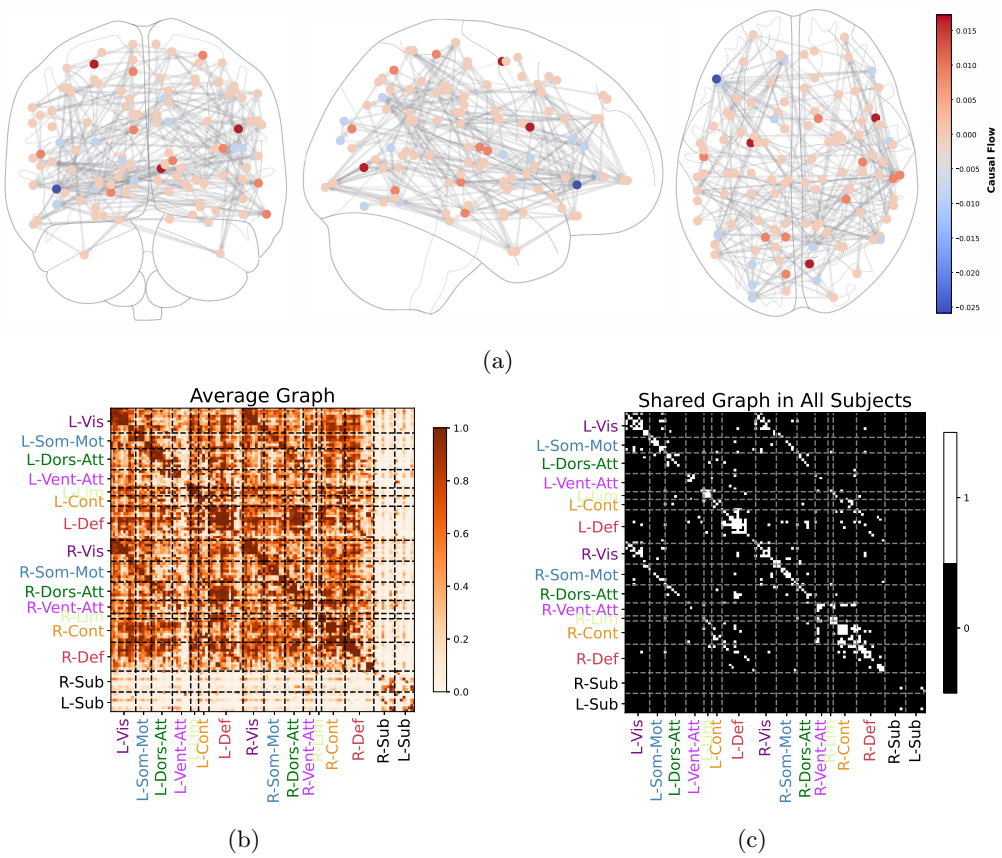


Figure 5: **Learned causal graphs are highly consistent across subjects.** (a) The subgraph that is shared among all 200 subjects (also shown in panel (c) in matrix form) as well as the corresponding causal flows for each node in this graph. (b) Average causal graph, computed as the mean of all the causal 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. Clear distinction can be seen in the causal links among cortical regions (colored labels) and subcortical ones (black labels). (c) 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.

shown in Figure 5a, the left prefrontal cortex (PFC) is a strong sink of causal effect while the left frontal eye field (FEF), right primary visual cortex (V1) and right ventral precentral cortex (PrCv) are strong sources. Such strong commonalities among subjects allow us to gain insights into the general patterns and characteristics of the causal relationships in a resting brain, even in the absence of ground truth data for direct comparison.

Net resting-state causal effect flows from ventral attention to visual network. Nodal centralities, such as the degree and causal flow (the sum and difference of out- and in-degrees, respectively), also show strong consistency among the subjects. Figure 6a shows the nodal degrees for all subjects (colored lines) as well as its mean across subjects (black line, also depicted in Figure 6b). Statistically significant differences exist between the degree distributions of many pairs of nodes (about 90% of the pairs have $p < 0.001$, 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 (see Supplementary Figures 28-31). 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.

Nodal causal flows also show strong similarities across subjects. While nodal degree quantifies the overall causal connectedness of a node, causal flow quantifies the asymmetry between in-flow and out-flow. Therefore, it is possible for a region to have a large degree but very small causal flow (balanced flow), or a relatively small degree but large causal flow (strong source or sink). As seen from Figure 6c,6d, bilateral medial ventral attention regions, e.g., are strong sources of causal flow, despite having notably low nodal degrees. Bilateral visual areas, on the other hand, are strong sinks of causal flow, though not particularly high in degree. Specific dorsal attention areas (namely, ventral precentral areas and frontal eye fields), as well as hippocampi, are also sources of causal flow bilaterally, even though neither showed strong nodal degrees. Therefore, in summary, we found bilateral strong causal flows emanating out of ventral attention, dorsal attention, and hippocampi, flowing through resting state network hubs in the default mode network (DMN) and thalami, and converging on the occipital lobe.

The strongest elements of this net causal flow can also be identified from an average subnetwork graph, as shown in Figure 6e (see Methods for detailed computations). Taking each subnetwork as a (hyper) node, we can also compute degrees and causal flows for each subnetwork, as shown in Figure 6f,6g (see Supplementary Figure 32 for nodal degrees of the average subnetwork graph and Supplementary Figure 33 for separate distributions of causal flow vs nodal degree for each subnetwork). We can therefore see that when averaged across all the regions within each subnetwork, bilateral ventral attention networks and the right limbic network act as sources of causal flow while bilateral visual networks and the left limbic network act as sinks. Therefore, we observe the strongest net causal flow during resting state to be from the ventral attention network to the visual network, as also seen more granularly from Figure 6a-6d.

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 7a shows the percentage of edges in the final graph which 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 34). Even further, such contemporaneous edges are substantially stronger than edges from lags 1-3 (Figure 7b). 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 weakly 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($r = -0.12, p = 10^{-43}$) (See Supplementary Figures 37 and 38 for separate maps of degree similarities and pairwise nodal distances). This relationship is stronger among intra-hemispheric parcels ($r = -0.27, p = 10^{-82}$) where connections are denser and shorter-distance, compared to inter-hemispheric parcels ($r = -0.09, p = 10^{-5}$). Thus, in summary, nodes that

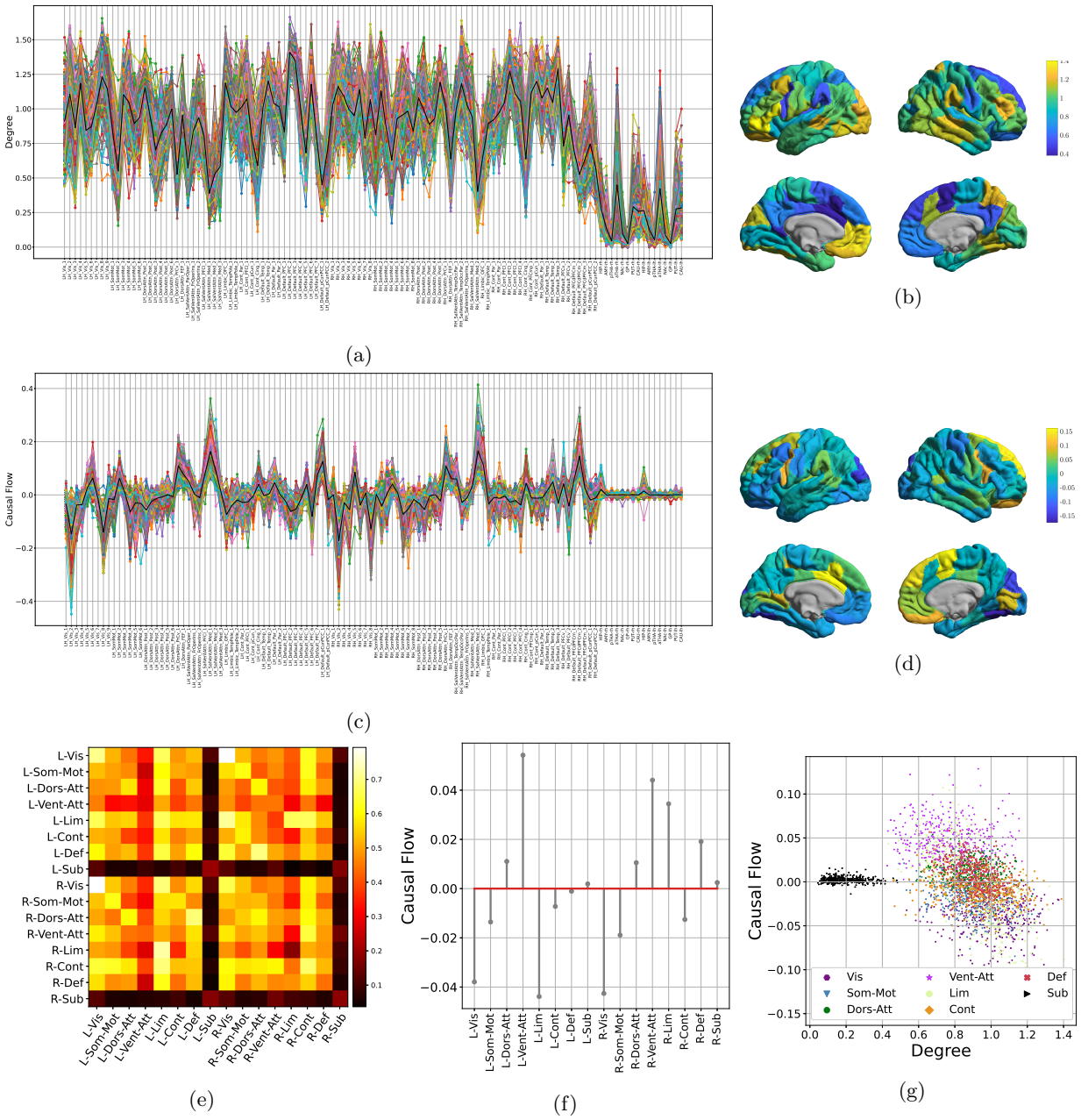
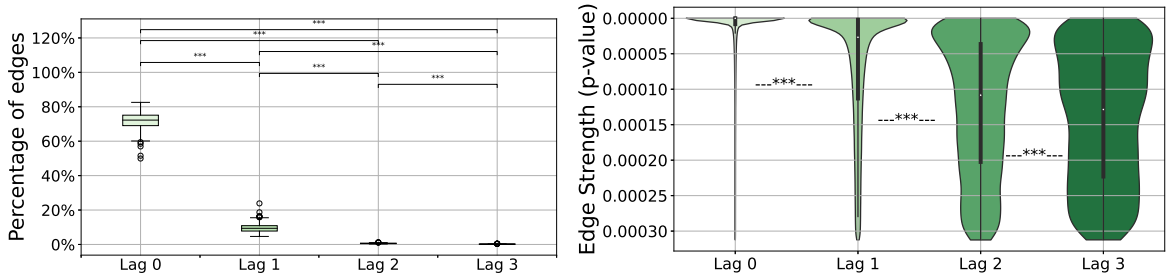


Figure 6: **Patterns of degree and causal flow across subjects.** (a) Distribution of nodal degree, computed separately for each node in the causal graph and each subject. Different colors correspond to different subjects and the black line shows the average nodal degrees across subjects. (b) Average cortical nodal degree (black line in (a)) shown over the brain cortical surface. (c,d) Similar to (a,b) but for nodal causal flows. (e) 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. Note that functional subnetworks are split over hemispheres for a more detailed analysis. (f) "Nodal" causal flows computed for the average subnetwork graph in (e). (g) The joint distributions of causal flow and degree for each "node" of the subnetwork graphs across all subjects.



(a) (b)

Figure 7: **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. (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. $^{***} = p < 0.001$, Wilcoxon signed-rank test.

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 39). 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 8b, 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 from Figure 8c, 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 9a, z = 14.61 and $p = 10^{-48}$, Wilcoxon signed-rank test), even though no such laterality was found in nodal causal flows (Figure 9b, z = 0.73 and $p = 0.23$, 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 9c shows the mean degrees of corresponding pairs of regions in the left and right hemispheres, color-coded by functional subnetworks (cf. Supplementary Figures 35 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. The similar plot for causal flows (Figure 9d, Supplementary Figure 36) 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. Similarly, we observe statistically significantly higher nodal degrees in women compared to men (Figure 9e, $p < 10^{-5}$, Wilcoxon

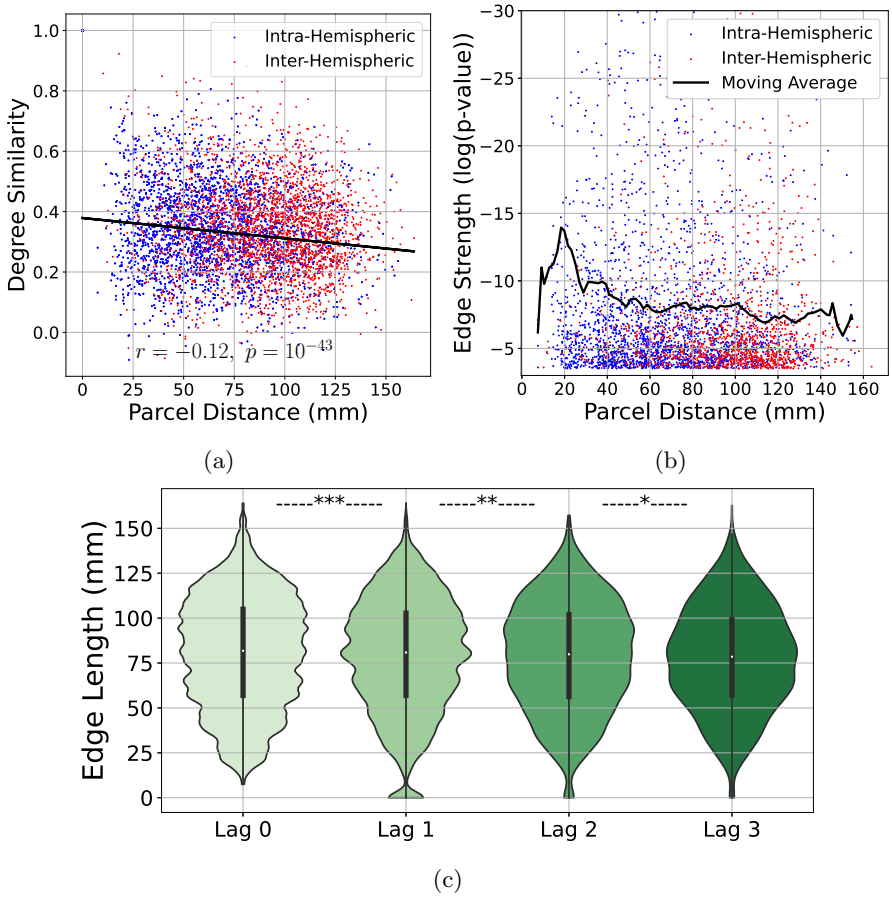


Figure 8: **Effect of Euclidean distance on edge attributes.** (a) Degree similarity (correlation coefficient between nodal degrees of two parcels over all subjects) as a function of the Euclidean distance between the parcels. Parcel pairs in the same hemisphere (intra-hemispheric) and parcels in two different hemispheres (inter-hemispheric) are shown in blue and red, respectively. Degree similarity decays statistically significantly with parcel distance ($r = -0.12, p = 10^{-43}$), much more so among intra-hemispheric parcels ($r = -0.27, p = 10^{-82}$) than inter-hemispheric ones ($r = -0.09, p = 10^{-5}$). (b) Edge strength (as in Figure 7b) as a function of Euclidean distance between edge end points (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. (c) Distributions of Euclidean edge length (parcel distance between edge endpoints) for edges existing in the subgraph corresponding to each lag. $^{*} = p < 0.05$, $^{**} = p < 0.01$, $^{***} = p < 0.001$, Wilcoxon signed-rank test.

signed-rank test), whereas a similar gender difference was not observed between causal flows (Figure 9f, $p = 0.81$, Wilcoxon signed-rank test).

Causal graphs are sparser and more consistent across subjects compared to functional connectivity. A major motivation for building causal connectomes is the removal of spurious connections in a functional connectivity (FC) profile 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)(Figure 10a,10b, no overlap existing between the support of the two distributions). In fact, FC graphs included approximately 95% of CaLLTiF's discovered causal edges (Figure 10c), while only about half of all functional connectivity edges are also causal (Figure 10d). 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 7a), but are fundamentally not discoverable by FC which only measures contemporaneous co-fluctuations. Moreover, we found causal connectomes significantly more consistent across subjects compared to FC connectomes (Figure 10e, $p < 0.001$, Wilcoxon signed-rank test), further reinforcing the expectation that causal edges are "pruned" and more reliable compared to functional edges.

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 shortcomings of existing methods, we proposed the Causal discovery for Large-scale Low-resolution Time-series with Feedback (CaLLTiF) algorithm that improves upon the state of the art in terms of computational efficiency, discovering contemporaneous effects, and correction for temporal multiple comparisons. We demonstrated the accuracy of CaLLTiF against several state-of-the-art approaches using simulated fMRI and then used it to estimate causal connectomes from resting-state HCP data [50, 66, 67]. On human fMRI data, CaLLTiF was able to learn causal graphs that (1) are highly consistent across subjects, (2) are consistent with known resting-state dynamics, (3) are sparser and more consistent across subjects compared to functional connectivity graphs, (4) properly reflect the presence of contemporaneous relationships due to the low temporal resolution of fMRI, (5) capture Euclidean distance-dependence in causal interactions, and (6) demonstrate statistically significant laterality and gender differences in degree distributions (but not causal flows). Overall, our results validate the power of CaLLTiF in detecting causal interactions from resting-state fMRI and open myriad avenues for future investigations of causal discovery from the task and stimulation-induced neuroimaging data.

As noted in Results, we used four criteria to narrow down the list of existing causal discovery methods based on their suitability for fMRI. First, we focused our analysis to algorithms that can learn graphs with cycles. This is due to the well-known prevalence and critical role of feedback connections in the brain [64, 65, 68]. Generally, two categories of algorithms can learn graphs with cycles: those that explicitly allow for cycles (which is rare), and algorithms that learn an (often acyclic) extended graphs over time-series data and cyclicity appears in their summary graph when collapsing the extended graph over time. We included algorithms from both categories, but found the former to generally have higher computational complexity and not scale well for whole-brain analysis. Second, neural signals are known to travel across the brain in much less than one fMRI repetition time (TR) [23, 24]. Such fast sub-TR interactions in fact account for a significant portion of neural dynamics and demonstrate themselves as "contemporaneous" effects in fMRI. Therefore, we further pruned algorithms based on their ability to learn contemporaneous effects. Third, we only selected algorithms that do *not* allow for latent nodes. This is due to a nuanced trade off between theoretical generality and practical utility. Allowing for latent nodes is in theory a capability and an advantage, but it comes at the cost of significantly expanding the set of graphs consistent with observational data. In turn, allowing for latent nodes *lowers* an algorithm's ability to detect unambiguous causal links. The possibility of whole-brain coverage and measuring all cortical and subcortical areas in fMRI provides a unique opportunity to avoid such costs and use simpler algorithms that assume a complete coverage of relevant variables in the observed data. Finally, we selected methods designed to capture linear

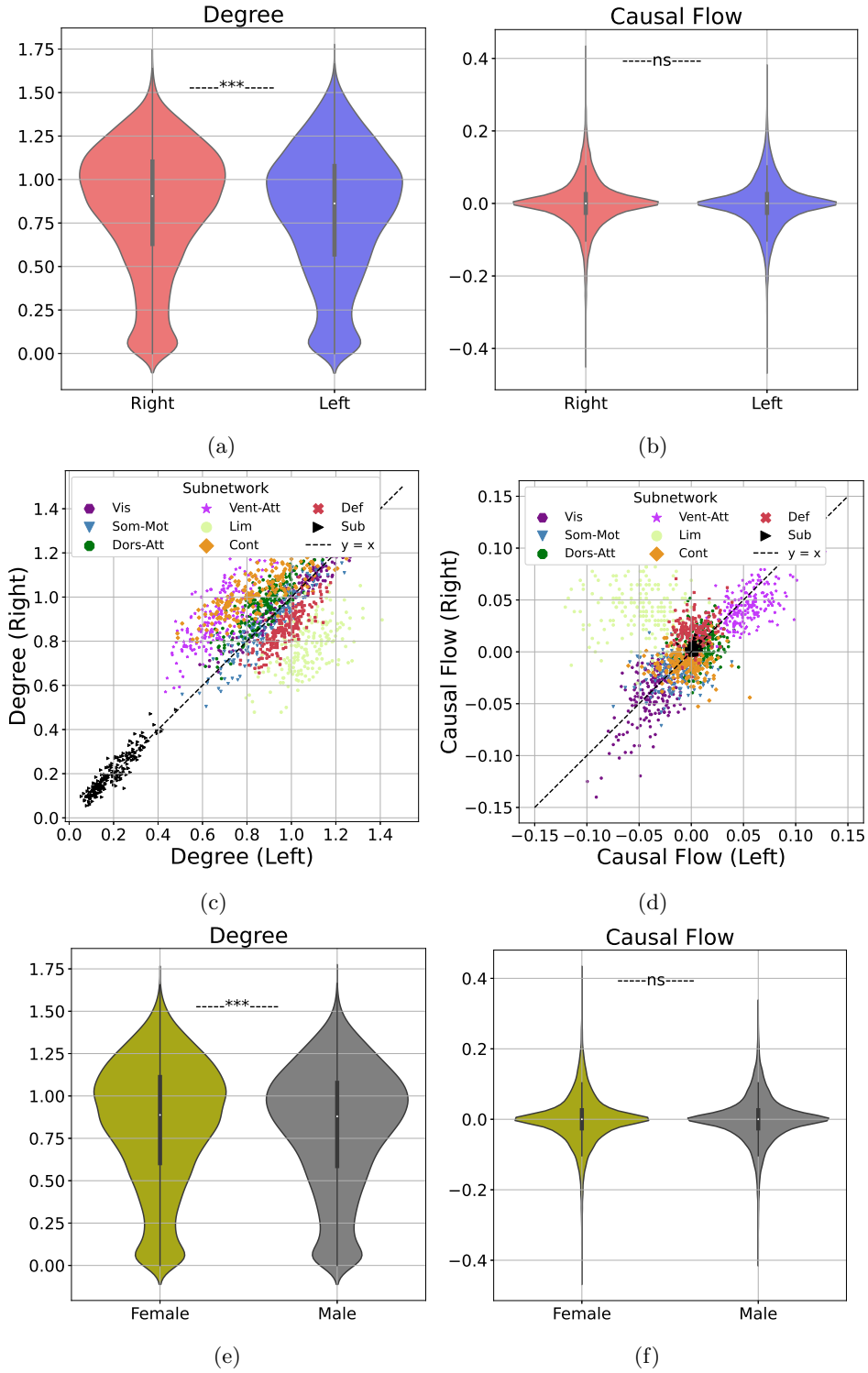


Figure 9: **Hemispheric laterality and gender differences in causal connectomes.** (a) Distributions of nodal degree in the right and left hemispheres, combined across all subjects. (b) Similar to (a) but for causal flows. (c) Nodal degrees, averaged across subjects and color-coded by functional subnetwork, 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.

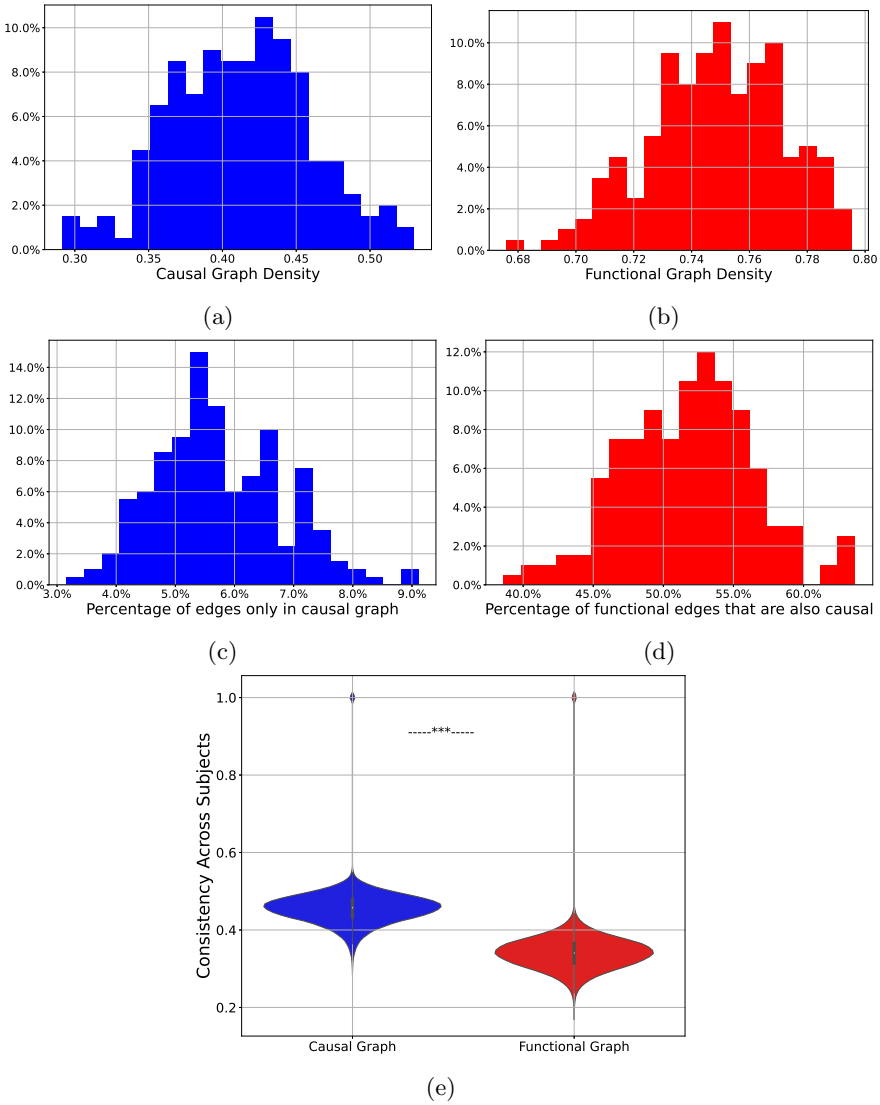


Figure 10: **Causal graphs are sparse and more consistent than functional connectivity (FC) graphs.** (a) Distribution of the density of edges in causal graphs learned by CaLLTiF. (b) Similar to (a) but for FC. (c) Distribution of the percentage of edges that only exist in CaLLTiF but not in FC. (d) Distribution of the percentage of edges that exist in both causal and functional graphs. (e) Distributions of graph consistencies (correlation coefficients) across subject. $^{***} = p < 0.001$, Wilcoxon signed-rank test.

causal relationships. This is mainly motivated by our recent empirical [69] and analytical [70, 71] works demonstrating the macroscopic linearity of resting-state dynamics, particularly those captured by fMRI. Nevertheless, we also tested a state-of-the-art nonlinear causal discovery method [14, 72] and found it to indeed have a lower F1 score compared to linear methods in synthetic fMRI (Supplementary Figure 22).

To address the limitations of existing algorithms, we proposed CaLLTiF which improves upon the state of the art in three main dimensions: type I error control, learning contemporaneous cycles, and scalability. In interpreting CaLLTiF's outputs, it is important to note the conservative nature of its method of correction for temporal multiple comparisons. For instance, 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. In other words, graphs returned by CaLLTiF are likely to have higher precision but lower recall than what is optimal based on F1 score. While such conservativeness is often desirable, it can also be tuned as needed by tuning the pre-correction significance threshold (q in Equation (4)).

A second core aspect of CaLLTiF is its treatment of contemporaneous effects. Our results with the HCP data (Figures 7a and 7b) 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 (Figure 8c) 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.

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. Figure 3). The approach taken by PCMCI increases statistical power (cf. the trend of optimal 'Alpha Level' values in Figure 3b), 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 24-26). 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 $X_i \rightarrow X_k \leftarrow X_j$). Finally, using PC-like methods for conditioning on carefully-chosen subsets of contemporaneous variables (as done, e.g., in PCMCI$^+$ [58]) requires an assumption of contemporaneous acyclicity and often increases computational complexity. Therefore, despite its heuristic nature, we found the conditioning in CaLLTiF to provide a desirable balance between precision and recall for the particular application of causal discovery from fMRI.

The third core aspect of CaLLTiF is its scalability. In the context of causal discovery, ROI-based analysis allows for a more targeted discovery among few variables of interest, makes it easier to validate obtained results against the available literature, and alleviates scalability concerns. Algorithms such as FASK [57] and DGlearn [62], e.g., have shown great promise when applied to small graphs. However, ROI-based analysis has two main limitations. The first limitation is an inherent bias towards already established networks and ROIs, limiting the power of causal 'discovery' to detect relationships not otherwise expected or hypothesized. The latter is increasingly possible using the power of emerging ever-larger datasets, but requires algorithmic solutions capable of handling big data efficiently. The second limitation is that ROI-based analysis mandates careful consideration of potential latent confounders. This significantly limits the range of algorithms one can use as well as the power and specificity of statistical methods in detecting unambiguous causal relationships in the presence of latent confounders. A significant need therefore exists for scalable algorithms such as CaLLTiF that can learn potentially cyclic, potentially contemporaneous causal relationships from whole-brain fMRI.

In the field of causal discovery, most methods assume that the underlying causal structure of a system can be represented by a Directed Acyclic Graph (DAG) [7, 36, 38–43]. In studying the brain, however, cycles and feedback are inevitable and pervasive [64, 65, 68, 73–75]. This significantly limits the range of methods that are suitable for causal discovery from neuroimaging data, primarily to two categories: time-series methods that still learn a DAG over lagged variables, and cross-sectional methods that allow for cycles even among contemporaneous variables (Table 1). The latter category deserves particular attention. Even

with a TR of 720ms, we found a strong majority of edges coming from contemporaneous effects. Moving to TRs of 2s or greater, which are still very common, will likely make contemporaneous effects almost all that matters. CaLLTiF provides a first step in developing scalable algorithms that can learn cycles also among contemporaneous variables, but only learns symmetric contemporaneous subgraphs. More research is therefore needed to design algorithms that can learn accurate, potentially asymmetric causal graphs among lagged and contemporaneous variables while scaling well to hundreds of variables.

Because the blood-oxygen-level-dependent (BOLD) signal measured by fMRI provides only an indirect measure of neural activity through the hemodynamic response function (HRF), one may first seek to estimate the underlying neural activations by deconvolving fMRI from the HRF and then use the estimated neural activations for causal discovery. We, however, did not find such deconvolution beneficial. Using the simulated Full Macaque data and standard Wiener deconvolution [76] with the generic SPM HRF [77], we compared the CaLLTiF, Mixed-PCMCI, and PCMCI methods when applied directly to the BOLD data or after HRF deconvolution (Supplementary Figure 23). While PCMCI and Mixed-PCMCI exhibited slightly higher F1 scores when applied to the deconvolved signals, the highest F1 scores were achieved by CaLLTiF over the BOLD signal. This is in line with our earlier findings regarding *decreased* dynamics signal to noise ratio after HRF deconvolution [69], and motivated using the BOLD signal itself in the rest of our analyses.

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 Figure 8c, 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. The lack of a ground-truth *causal* connectome further makes the validation of obtained graphs challenging. Unlike structural connectivity, a causal connectome is directed, dynamic, and task-specific. As such, once an algorithm is sufficiently validated on simulated data, its outcomes on real data must be 'trusted' to some extent. Our conservative method for correction for temporal multiple comparisons seeks to serve this purpose and ensure a certain level of statistical confidence in each edge of the resulting graphs. Finally, 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 [78].

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. Structural connectivity can also be used as a prior for causal discovery and augment functional data. Complementary to our focus on learning separate causal graphs for each subject, one can also learn one causal graph from a group of subjects and use that group-level graph to improve the estimation of each causal graph [79, 80]. Multimodal discovery, combining, e.g., fMRI and EEG, is also highly warranted given the complementary spatial and temporal resolutions of different imaging modalities. We leave these exciting possibilities to future studies.

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 [57]. 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 [57]. In brief, the same simulation procedure was used for simple and macaque-based graphs, where the authors used the model proposed in [45] which is itself based on the DCM architecture of [26]. 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 [45, 81] 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 [50, 66, 67]. 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 [82] 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 [83]) and 16 subcortical ones (Melbourne Scale I atlas [84]).

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 [1] 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 [38]. 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 [1]. However, PCMCI is considered not suitable for highly predictable (almost deterministic) systems with little new information at each time step [1]. 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' (cf. Figure 4), 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 [58]. 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 [1] 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 [59], 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 [60], 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 [85], 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 [62], 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 [57], 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 [57] and Java source code for it is available at `http://github.com/cmu-phil/tetrad`.

MVGC. In [51], 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 [8]. While this allows for optimal forecasting of an effect and has been extended to multivariate systems [20], 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 [72]. 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 [85], 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 .22). 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 [1] 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 Figure 3, 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 ($X_i(t-\tau) \rightarrow X_j(t)$ 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 includes 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 compatible with the lagged direction(s) otherwise. 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

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

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 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 1, 2,3,5,7,8,10 for the simulated Simple Network graphs, in Supplementary Figures 11, 12,13,14,15 for the simulated Small-Degree Macaque data, and in Supplementary Figures 16, 17, 18, 19, 20, 21, and 22 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 [69], 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 7a). 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 [86–90]. 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 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 [91] plus one subcortical subnetwork, separately for 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 subject, as depicted in Figure 6e.

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 and by the Canadian Institute for Advanced Research (fellowship awarded 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

[1] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets, Science Advances 5 (11) (2019) eaau4996. doi:10.1126/sciadv.aau4996.

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

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

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

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

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

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

[8] C. K. Assaad, E. Devijver, E. Gaussier, Survey and evaluation of causal discovery methods for time series 73 (2022). doi:10.1613/jair.1.13428.

[9] J. B. Goense, N. K. Logothetis, Neurophysiology of the bold fmri signal in awake monkeys, Current Biology 18 (9) (2008) 631–640. doi:https://doi.org/10.1016/j.cub.2008.03.054.

[10] Z. Q. Winder AT, Echagarruga C, D. PJ, Weak correlations between hemodynamic signals and ongoing neural activity during the resting state, Nature Neuroscience 18 (2017) 17611769–640. doi:10.1038/s41593-017-0007-y.

[11] S. L. Bressler, V. Menon, Large-scale brain networks in cognition: emerging methods and principles, Trends in Cognitive Sciences 14 (6) (2010) 277–290. doi:https://doi.org/10.1016/j.tics.2010.04.004.
URL https://www.sciencedirect.com/science/article/pii/S1364661310000896

[12] E. Bullmore, O. Sporns, The economy of brain network organization, Nature Reviews Neuroscience 13 (5) (2012) 336–349.

[13] A. Fornito, E. T. Bullmore, Connectomics: A new paradigm for understanding brain disease, European Neuropsychopharmacology 25 (5) (2015) 733–748, advances in neuroscience for improved mental health: an ECNP analysis of progress and prospects. doi:https://doi.org/10.1016/j.euroneuro.2014.02.011.

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

[15] A. Gerhardus, J. Runge, High-recall causal discovery for autocorrelated time series with latent confounders, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 12615–12625.

[16] D. Malinsky, P. Spirtes, Causal structure learning from multivariate time series in settings with unmeasured confounding, in: T. D. Le, K. Zhang, E. Kıcıman, A. Hyvärinen, L. Liu (Eds.), Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, Vol. 92 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 23–47.

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

[18] J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, C. Glymour, Six problems for causal inference from fmri, Neuro]mage 223 (2010) 1545–1558. doi:https://doi.org/10.1016/j.neuroimage.2009.08.065.

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

[20] L. Barnett, A. K. Seth, The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference, Journal of Neuroscience Methods 223 (2014) 50–68. doi:https://doi.org/10.1016/j.jneumeth.2013.10.018.

[21] S. A. Huettel, A. W. Song, G. McCarthy, Functional Magnetic Resonance Imaging, 2nd Edition, Sinauer Associates, 2009.

[22] P. L. Nunez, R. Srinivasan, Electric Fields of the Brain: The Neurophysics of EEG, Oxford University Press, 2006.

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

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

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

[26] K. Friston, L. Harrison, W. Penny, Dynamic causal modelling, NeuroImage 19 (4) (2003) 1273–1302. doi:https://doi.org/10.1016/S1053-8119(03)00202-7.

[27] K. J. Friston, J. Kahan, B. Biswal, A. Razi, A dcm for resting state fmri, NeuroImage 94 (2014) 396–407. doi:https://doi.org/10.1016/j.neuroimage.2013.12.009.
URL https://www.sciencedirect.com/science/article/pii/S1053811913012135

[28] K. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, P. Zeidman, Dynamic causal modelling revisited, NeuroImage 199 (2019) 730–744. doi:https://doi.org/10.1016/j.neuroimage.2017.02.045.

[29] K. J. Friston, J. Kahan, B. Biswal, A. Razi, A dcm for resting state fmri, NeuroImage 94 (2014) 396–407. doi:https://doi.org/10.1016/j.neuroimage.2013.12.009.

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

[31] Analysing connectivity with granger causality and dynamic causal modelling, Current Opinion in Neurobiology 23 (2) (2013) 172–178, macrocircuits. doi:https://doi.org/10.1016/j.conb.2012.11.010.

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

[33] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2009.

[34] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan kaufmann, 1988.

[35] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, MIT press, 2000.

[36] J. Pearl, Causal inference in statistics: An overview (2009).

[37] P. Spirtes, K. Zhang, Causal discovery and inference: concepts and recent methodological advances, in: Applied informatics, Vol. 3, SpringerOpen, 2016, pp. 1–28.

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

[39] P. Spirtes, C. Meek, T. Richardson, Causal inference in the presence of latent variables and selection bias, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 499–506.

[40] C. Meek, Graphical models: selecting causal and statistical models, Ph.D. thesis, Carnegie Mellon University (1997).

[41] C. Meek, Causal inference and causal explanation with background knowledge, in: Conference on Uncertainty in Artificial Intelligence, 1995.

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

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

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

[45] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, M. W. Woolrich, Network modelling methods for fmri, NeuroImage (2011) 875–891doi:https://doi.org/10.1016/j.neuroimage.2010.08.063.

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

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

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

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

[50] D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, The wu-minn human connectome project: an overview, NeuroImage 80 (2013) 62–79. doi:doi:10.1016/j.neuroimage.2013.05.041.

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

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

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

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

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

[56] J. Ramsey, K. Zhang, M. Glymour, R. S. Romero, B. Huang, Immé, Ebert-Uphoff, S. M. Samarasinghe, E. A. Barnes, C. Glymour, Tetrad - a toolbox for causal discovery, 2018.

[57] R. Sanchez-Romero, J. Ramsey, K. Zhang, M. Glymour, B. Huang, C. Glymour, Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods, Network Neuroscience 86 (2019) 274–306. doi:doi:10.1162/netn_a_00061.

[58] J. Runge, Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 1388–1397.

[59] A. Hyvärinen, K. Zhang, S. Shimizu, P. O. Hoyer, Estimation of a structural vector autoregression model using non-gaussianity, Journal of Machine Learning Research 11 (56) (2010) 1709–1731. URL http://jmlr.org/papers/v11/hyvarinen10a.html

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

[61] G. Lacerda, P. Spirtes, J. Ramsey, P. Hoyer, Discovering cyclic causal models by independent components analysis, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, 2008.

[62] A. Ghassami, A. Yang, N. Kiyavash, K. Zhang, Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs, in: International Conference on Machine Learning, 2020.

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

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

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

[66] D. M. Barch, Resting-state functional connectivity in the human connectome project: Current status and relevance to understanding psychopathology, Harvard Review of Psychiatry 25 (5) (2017) 209–217. doi:10.1097/HRP.0000000000000166.

[67] G. Burgess, S. Kandala, D. Nolan, T. Laumann, J. Power, B. Adeyemo, M. Harms, S. Petersen, D. Barch, 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) (2016) 669–680. doi:10.1089/brain.2016.0435.

[68] N. T. Markov, M. Ercsey-Ravasz, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, H. Kennedy, Cortical high-density counterstream architectures, Science 342 (6158) (2013) 1238406.

[69] E. Nozari, J. Stiso, L. Caciagli, E. J. Cornblath, X. He, M. A. Bertolero, A. S. Mahadevan, G. J. Pappas, D. S. Bassett, Is the brain macroscopically linear? a system identification of resting state dynamics, Nature Biomedical EngineeringIn press, available at https://www.biorxiv.org/content/early/2020/12/22/2020.12.21.423856 (2023).

[70] S. Ahmed, E. Nozari, On the linearizing effect of spatial averaging in large-scale populations of homogeneous nonlinear systems, in: 2022 IEEE 61st Conference on Decision and Control (CDC), IEEE, 2022, pp. 641–648.

[71] S. Ahmed, E. Nozari, On the linearizing effect of temporal averaging in nonlinear dynamical systems, in: Proceedings of the American Control Conference (ACC), IEEE, 2023, in press.

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

[73] K. J. Friston, B. Li, J. Daunizeau, K. E. Stephan, Network discovery with dcm, Neuroimage 56 (3) (2011) 1202–1221.

[74] R. Kötter, K. E. Stephan, Network participation indices: characterizing component roles for information processing in neural networks, Neural Networks 16 (9) (2003) 1261–1275.

[75] S. Zeki, S. Shipp, The functional logic of cortical connections, Nature 335 (6188) (1988) 311–317.

[76] N. Wiener, Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications, MIT press, 1949.

[77] K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. Rugg, R. Turner, Event-related fmri: characterizing differential responses, Neuroimage 7 (1) (1998) 30–40.

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

[79] K. M. Gates, P. C. Molenaar, Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples, NeuroImage 63 (1) (2012) 310–319.

[80] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, B. Schölkopf, Causal discovery from heterogeneous/nonstationary data, The Journal of Machine Learning Research 21 (1) (2020) 3482–3534.

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

[82] J. Elam, 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 (May 2020).

[83] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, Cerebral cortex 28 (9) (2018) 3095–3114. `doi:10.1093/cercor/bhx179`.

[84] Y. Tian, D. S. Margulies, M. Breakspear, A. Zalesky, Hierarchical organization of the human subcortex unveiled with functional connectivity gradients, bioRxiv (2020). `doi:10.1101/2020.01.22.916357`.

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

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

[87] A. Tharwat, Classification assessment methods, Applied Computing and Informatics 17 (1) (2020) 168–192.

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

[89] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233–240.

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

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