A Comparison with existing methods

In this section, we compared our proposed model with 4 existing methods, which were the the PC algorithm [1], GES [2], ICA-LiNGAM [3] and direct-LiNGAM [4] through a series of simulation studies. The graph generations were the same as described in the main paper; see Section III.A. We simulated the random DAG G through the R package *pcalg* with the edge probability $d/(p-1)$, where d is an edge degree parameter and p is the total number of variables. Given G, we assigned uniformly random weights to the edges to obtain the weighted adjacency matrix $\mathbf{B}$: $b_{ij} \sim \text{Unif}(-0.8, -0.3) \cup (0.3, 0.8)$, if $b_{ij} \in E$, otherwise $b_{ij} = 0$. Given $\mathbf{B}$, we generated $\mathbf{x} = \mathbf{B}^T\mathbf{x} + \boldsymbol{\epsilon} \in \mathbb{R}^p$ from two non-Gaussian noise selections: Exponential (Exp) and Chi-squared (Chisq) noise. The exponential noise was set to have rate 1, i.e., $\epsilon_i \sim exp(1)$, $i = 1, 2, ..., p$, and the chi-squared noise was set to be central with a degree of freedom 1, i.e., $\epsilon_i \sim \chi^2_1$, $i = 1, 2, ..., p$. We then sampled the random vectors $\mathbf{x} \in R^{n\times p}$ for each noise selection with $n = 500$, $p = 50, 100, 200$ and the degree parameter $d = 1, 2, 4$ based on the models. For each scenario, we simulated 10 datasets independently. All these algorithms can be implemented through the R package *pcalg*. We set the significance level $\alpha = 0.05$ (with FDR correction) for all 5 methods to obtain the estimated DAGs.

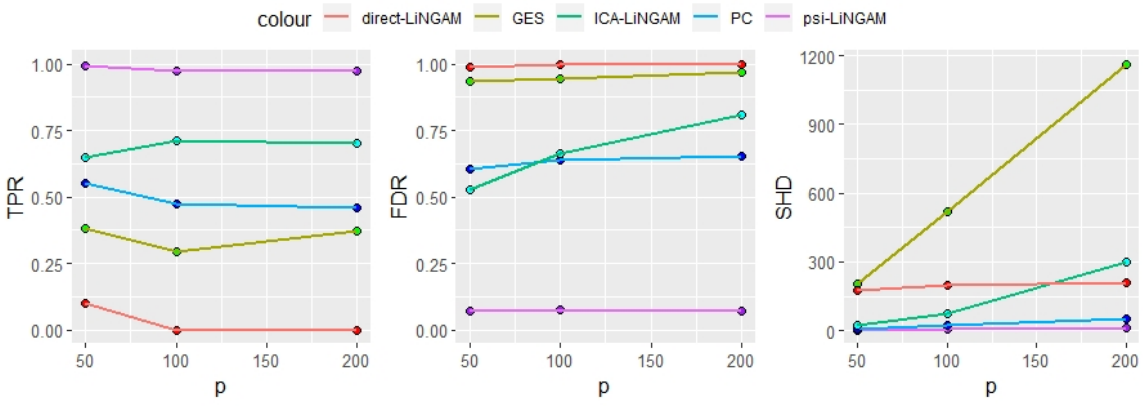
We assessed the performance of the 5 methods through the true positive rate (TPR), false discovery rate (FDR), and structural hamming distance (SHD) [5]. SHD is a commonly used metric based on the number of operations needed to transform the estimated DAG into the true graph [6]. In simple terms, SHD counts the total number of edge insertions, deletions or flips during the transformation. TPR and FDR are two typical measures of a binary classification. Let us define an experiment from P positive instances and N negative instances for some conditions. In our case, the positive instance represents a directed edge from one node to the other. The four outcomes are summarized in Table 1. The definitions of TPR and FDR are given as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}}.$$

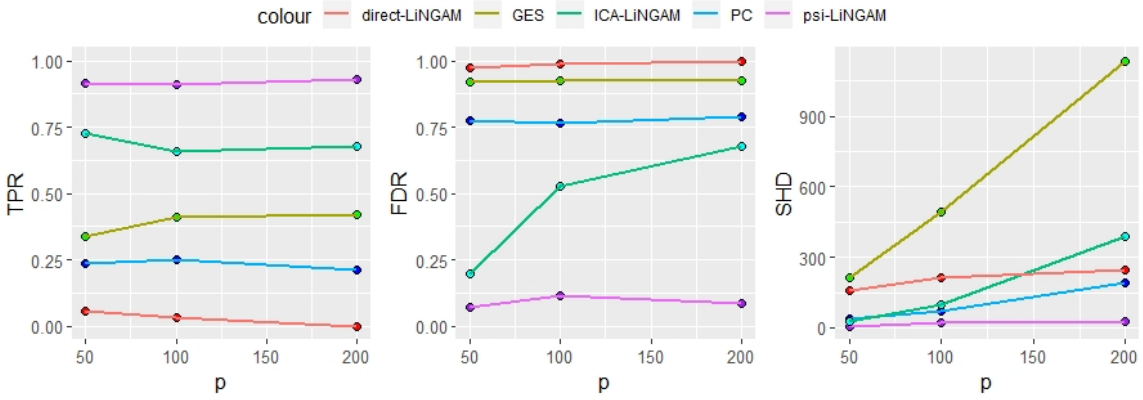
Table 1: Outcomes of a binary decision

	Actual positive (P)	Actual negative (N)
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

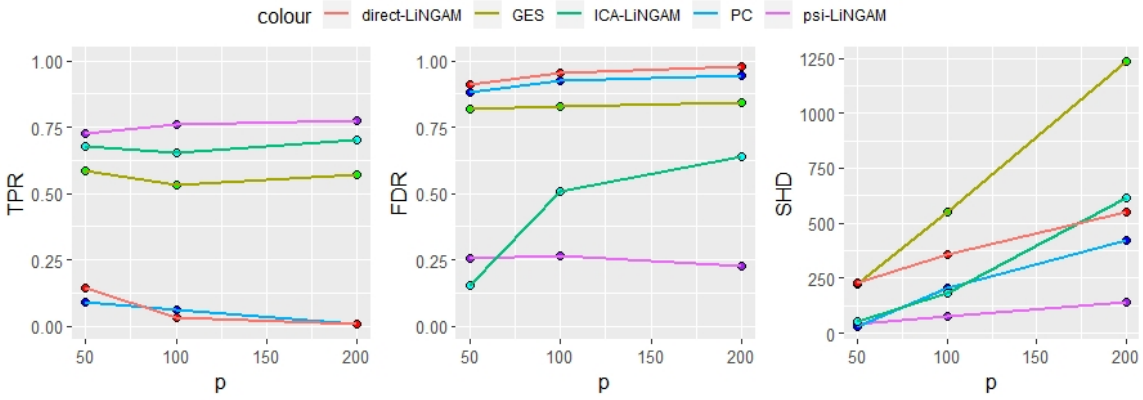
The simulation results are shown in Fig. A and Fig. B. As we can see, the ψ-LiNGAM has significantly improved performance over direct-LiNGAM. The ψ-LiNGAM has the highest TPR while maintaining a low range of FDR under each setting. Although the SHD grows with the variable size p, the increasing slope of the ψ-LiNGAM is the lowest as well as the SHD value. Overall, our proposed ψ-LiNGAM has outperformed the other methods under each setting, especially under large variable number and/or low degree parameter setting. For PC algorithm and GES, the low TPR and high FDR are caused by their poor direction identification. To be more specific, both PC and GES can only estimate the completed partially directed acyclic graph (CPDAG), which contains both undirected and directed edges [1, 2]. For the undirected edges in PC and GES, we treat them as bi-directional to calculate the metrics. The high SHD value of GES is because the GES method tends to identify more false positive edges. Although LiNGAM has a decent performance compared to PC and GES, the false discovery rate has increased significantly as the variable size increases. This is because the original LiNGAM method needs a large number of samples in the relevant dimension to converge.



(a) d=1

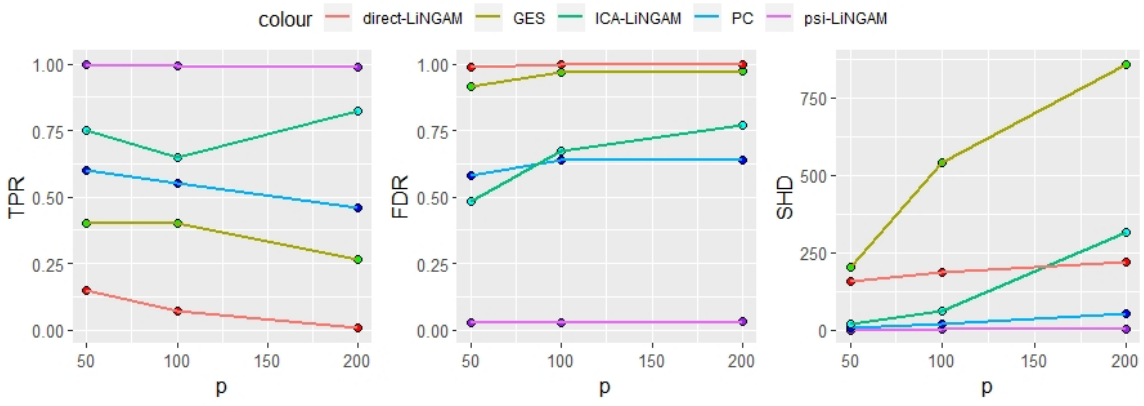


(b) d=2

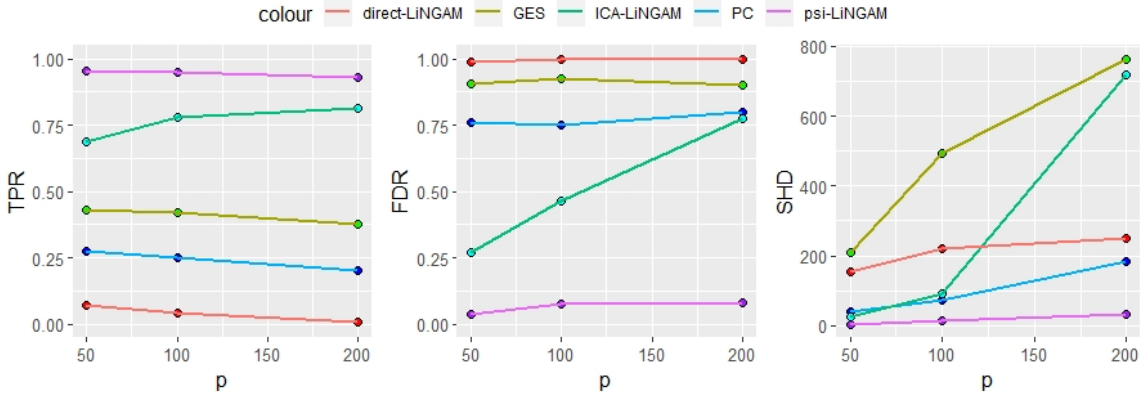


(c) d=4

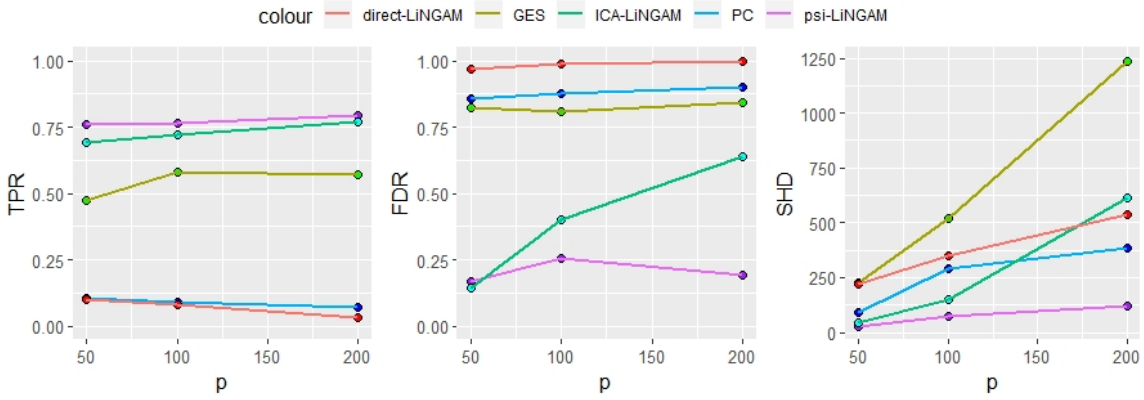
Figure A: Simulation results with the exponential noise setting, which represent the average performance in terms of TPR, FDR and SHD under various variable ($p = 50, 100, 150$) and degree parameter ($d = 1, 2, 4$) settings with $n = 500$.



(a) d=1



(b) d=2



(c) d=4

Figure B: Simulation results with chi-squared noise setting, which represent the average performance in terms of TPR, FDR and SHD under various variable ($p = 50, 100, 150$) and degree parameter ($d = 1, 2, 4$) settings with $n = 500$.

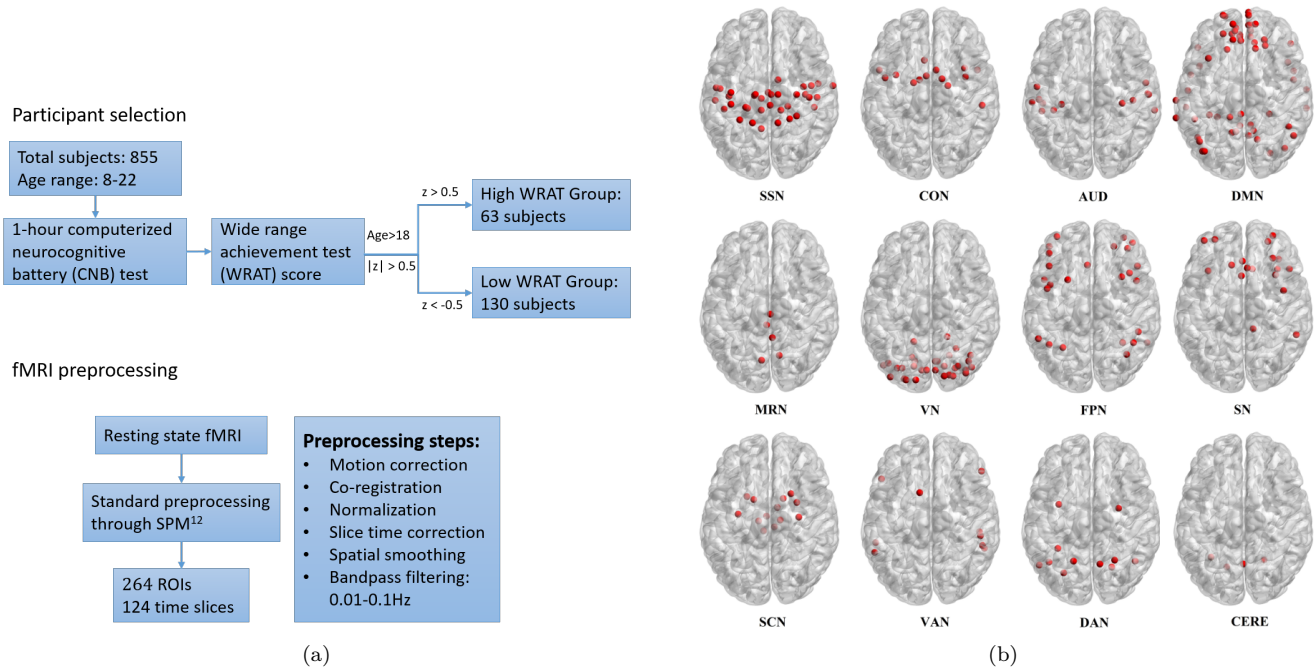
B Computation

Compared with the direct-LiNGAM, the computational complexity of the ψ-LiNGAM is favorable. We recorded the mean CPU times of the two methods based on the simulation studies in Section A with various variable sizes p's and graph densities ($d = 1, 2, 4$). The simulations were run on a 4 GHz computer. The results are shown in Table 2.

Table 2: The mean CPU times (in seconds) for $d = 1, 2, 4$ under various variable sizes p's.

Method	$p = 50$	$p = 100$	$p = 200$
ψ-LiNGAM	28.86	291.40	2130.26
Direct-LiNGAM	51.83	499.00	3806.14

C Supplementaries of the fMRI studies



(a) (b)

Figure C: (a) Illustration of the data selection and preprocessing. (b) 12 functional networks considered in this paper. These networks are expressed in the 264 nodes of the template defined by Power et al. ([7]).

Table 3: The 16 causal connections selected for $|d| > 0.4$, where L and R represent the left and the right side of the brain, respectively.

ROI					ROI			
Ind.	MNI	AR (L/R)	FN		Ind.	MNI	AR (L/R)	FN
85	$(27, 16, -17)$	Insula (R)	-	$\longrightarrow$	82	$(46, 16, -30)$	Temporal mid pole (R)	DMN
[1,2,4] 137	$(-46, 31, -13)$	Inferior frontal gyrus, orbital (L)	DMN	$\longrightarrow$				
[1−4] 87	$(-39, -75, 44)$	Angular gyrus (L)	DMN	$\longrightarrow$	134	$(-7, -71, 42)$	Precuneus (L)	MRN
96	$(52, -59, 36)$	Angular gyrus (R)	DMN	$\longrightarrow$	250	$(-50, -7, -39)$	Inferior temporal gyrus (L)	-
122	$(12, 36, 20)$	Anterior cingulate gyrus (R)	DMN	$\longrightarrow$	242	$(-49, 25, -1)$	Inferior frontal gyrus, triangular (L)	VAN
212	$(-11, 26, 25)$	Anterior cingulate gyrus (L)	SN	$\longrightarrow$				
[1] 128	$(52, 7, -30)$	Temporal mid pole (R)	DMN	$\longrightarrow$	81	$(-44, 12, -34)$	Temporal mid pole (L)	DMN
[1] 163	$(6, -72, 24)$	Cuneus (R)	VN	$\longrightarrow$	1	$(-25, -98, -12)$	Inferior occipital lobe (L)	-
[1] 165	$(26, -79, -16)$	Fusiform gyrus (R)	VN	$\longrightarrow$	158	$(20, -86, -2)$	Lingual gyrus (R)	VN
[1] 214	$(-28, 52, 21)$	Middle frontal gyrus (L)	SN	$\longrightarrow$	107	$(-7, 51, -1)$	Anterior cingulate gyrus (L)	DMN
[1] 218	$(31, 56, 14)$	Middle frontal gyrus (R)	SN	$\longrightarrow$	215	$(0, 30, 27)$	Anterior cingulate gyrus (L)	SN
[1] 224	$(-10, -18, 7)$	Thalamus (L)	SCN	$\longrightarrow$	225	$(12, -17, 8)$	Thalamus (R)	SCN
[1,2,4] 230	$(23, 10, 1)$	Putamen (R)	SCN	$\longrightarrow$	231	$(29, 1, 4)$	Putamen (R)	SCN
243	$(-16, -65, -20)$	Cerebellum 6 (L)	CERE	$\longrightarrow$	246	$(1, -62, -18)$	Vermis 6 (R)	CERE
[1,3,4] 251	$(10, -62, 61)$	Precuneus (R)	DAN	$\longrightarrow$	256	$(22, -65, 48)$	Superior parietal gyrus (R)	DAN
259	$(-33, -46, 47)$	Inferior parietal gyrus (L)	DAN	$\longrightarrow$	92	$(8, -48, 31)$	Anterior cingulate gyrus (R)	DMN

[1] There is a significant WRAT effect on the connection.

[2] There is a significant gender effect on the connection.

[3] There is a significant age effect on the connection.

[4] There is a significant age × gender interaction on the connection.

References

[1] P. Spirtes, C.N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

[2] D.M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3:507–554, 2002.

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

[4] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P.O. Hoyer, and K. Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

[5] I. Tsamardinos, L.E. Brown, and C.F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

[6] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[7] J. D. Power, D. A. Fair, B. L. Schlaggar, and S. E. Petersen. The development of human functional brain networks. *Neuron*, 67(5):735–748, 2010.