Journal of Machine Learning for Biomedical Imaging 2024:028 vol. 2, pp. 2223–2257 Submitted 05/2024
Published 11/2024

GLACIAL: Granger and Learning-based Causality Analysis for Longitudinal Imaging Studies

Minh Nguyen https://orcid.org/0000-0003-4762-1798 bn244@cornell.edu
ECE Department, Cornell Tech, New York, NY, USA

Gia H. Ngo https://orcid.org/0000-0002-2793-2700 ghn8@cornell.edu
ECE Department, Cornell Tech, New York, NY, USA

Mert R. Sabuncu https://orcid.org/0000-0002-7068-719X msabuncu@cornell.edu
ECE Department, Cornell Tech, New York, NY, USA

for the Alzheimer's Disease Neuroimaging Initiative*

Abstract

The Granger framework is useful for discovering causal relations in time-varying signals. However, most Granger causality (GC) methods are developed for densely sampled time-series data. A substantially different setting, particularly common in medical imaging, is the longitudinal study design, where *multiple* subjects are followed and sparsely observed over time. Longitudinal studies commonly track several biomarkers, which are likely governed by nonlinear dynamics that might have subject-specific idiosyncrasies and exhibit both direct and indirect causes. Furthermore, real-world longitudinal data often suffer from widespread missingness. GC methods are not well-suited to handle these issues. In this paper, we propose an approach named GLACIAL (Granger and LeArning-based CausalIty Analysis for Longitudinal studies) to fill this methodological gap by marrying GC with a multi-task neural forecasting model. GLACIAL treats subjects as independent samples and uses the model's average prediction accuracy on hold-out subjects to probe causal links. Input dropout and model interpolation are used to efficiently learn nonlinear dynamic relationships between a large number of variables and to handle missing values respectively. Extensive simulations and experiments on a real longitudinal medical imaging dataset show GLACIAL beating competitive baselines and confirm its utility. Our code is available at `https://github.com/mnhng/GLACIAL`.

Keywords: Machine Learning, Medical Imaging, Longitudinal Studies, Causality

1. Introduction

Granger causality (GC) (Granger, 1969) is a versatile and popular framework that exploits "the arrow of time" to detect causal relations in timeseries data (Roebroeck et al., 2005; Zhang et al., 2011). In GC, we test whether past values of one time series predict the future values of another series (i.e., forecasting), which allows us to infer causal relationships. Despite its popularity, current implementations of GC are only well-suited for densely and

∗. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

https://doi.org/10.59275/j.melba.2024-c7eg

uniformly sampled timeseries data from one system at a time. They are not designed for the longitudinal setup involving multiple systems (e.g., subjects), which are common in medical imaging. Although one could infer a causal graph for each subject and aggregate the graphs across subjects, this approach is untenable in many longitudinal studies in medicine where each subject only has a few observations, making the inference of each causal graph inaccurate.

Constraint-based methods such as PC or FCI (Spirtes et al., 2000), which rely on independent samples and conditional independence tests, are also commonly used for causal discovery. These methods would use one observation per subject and thus are not designed to detect causal relations reflected in temporal dynamics. We believe there is a lack of methods for causal discovery in longitudinal studies that consist of multiple subjects with sparse observations.

Longitudinal imaging studies typically track several variables simultaneously. Thus, applying causal discovery to longitudinal studies can be challenging because of the large number of variables involved and the complex (nonlinear) relationships between variables. Nonlinear GC methods (e.g. those based on non-parametric methods (Su and White, 2007; Marinazzo et al., 2008)) do not scale to large number of variables (Eichler, 2012). Similarly, existing GC tests that use neural networks to infer nonlinear dynamics (Tank et al., 2021; Nauta et al., 2019; Khanna and Tan, 2020) also face scalability issues. On the other hand, using linear GC to infer nonlinear relationships can be fast but may produce wrong results (Li et al., 2018).

Furthermore, prior GC methods are, to the best of our knowledge, all association-based. That is, they test for causal relationships via interrogating fit (learned) model weights. For example, in the linear GC approaches, this is achieved by testing the statistical significance of model coefficients. As the detection power of association-based GC (Granger, 1969; Lütkepohl, 2005) diminishes with increasing number of variables (Sugihara et al., 2012; Runge et al., 2019b), it may fail to detect the weak coupling between a node and its parents, in particular when there are a lot of variables and limited data (Runge et al., 2019a; Yuan and Shou, 2022). Another challenge of real-world longitudinal studies is missing data. While there is no consensus about what to do about missing values (Glymour et al., 2019), several works (Strobl et al., 2018; Tu et al., 2019) have tried to address this issue for cross-sectional data. Yet, as far as we know, missingness is under-explored in longitudinal studies, particularly in the context of causal discovery. Finally, GC, in its original form, does not differentiate between direct and indirect causes (Yuan and Shou, 2022). Although, in theory, infinite history (observations) could shield off indirect causes from being detected as edges in the output causal graph, when the number of observations per subject is small, false positives due to indirect causes is a common practical problem.

In this work, we propose GLACIAL (Fig 1), which stands for a "Granger and LeArning-based CausalIty Analysis for Longitudinal studies." GLACIAL combines GC with a practical machine-learning based approach to test for causal relations among multiple variables in a longitudinal study. GLACIAL extends GC to longitudinal studies by treating each subject's trajectory as an independent sample, governed by a shared causal mechanism that is reflected in the temporal dynamics. This treatment is similar to prior works where subjects are assumed to be independent samples in longitudinal data analysis (Hernan and Robins, 2020). By applying a standard train-test setup with hold-out subjects, GLACIAL can test

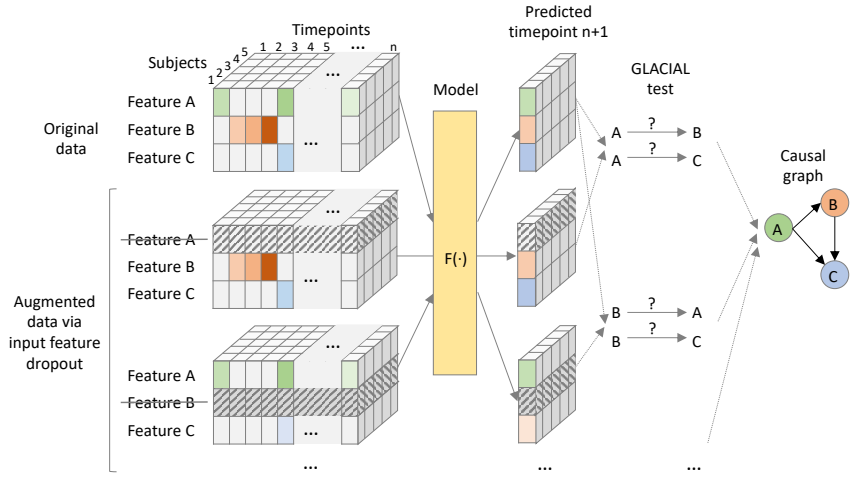


Figure 1: GLACIAL. Overview of the proposed approach for longitudinal studies.

for effects of causal relations in expectation. Critically, GLACIAL infers causal relationships based on interrogating predictive accuracy and not a direct analysis of model weights, which is common in existing association-based GC methods. GLACIAL employs a single multi-task neural forecasting model, trained with input feature drop-out, to learn nonlinear relationships among all variables in time-varying data. The model also handles missing values using model interpolation. Thus, although neural networks have been used in the past for causal discovery, GLACIAL efficiently tests for causal relations of a large set of variables in data where timepoints may be sampled irregularly and may contain missing values. The efficiency and flexibility of GLACIAL make it applicable to real-world multimodal medical imaging studies with many variables. Furthermore, GLACIAL includes post-processing heuristics to account for indirect causes and resolve the directionality of detected ambiguous associations. Extensive experiments on synthetic data and real data from a longitudinal medical imaging study show that GLACIAL can infer relationships accurately even in challenging real-world scenarios with sparse observations, a large number of variables and direct causes, and a large degree of missing data. Although a specific model was used in our experiments, GLACIAL is model-agnostic.

2. Related Works

Most existing causal discovery (CD) methods are not intended for the longitudinal study design, where multiple subjects are sparsely observed at different timepoints. CD methods designed for timeseries data or independent samples are often used in the longitudinal setting despite potential poor performance.

Causal Discovery: CD methods intended for cross-sectional studies are ill-suited for longitudinal studies. They often fall under: constraint-based search (e.g. FCI (Spirtes et al., 2000)), score-based search (e.g. Greedy Equivalence Search (GES) (Chickering, 2002)), functional causal models (FCMs) (Shimizu et al., 2006; Hoyer et al., 2008; Zhang and

Hyvärinen, 2009a; Zhang and Chan, 2006; Zhang and Hyvärinen, 2009b), or continuous optimization (Zheng et al., 2018). Search methods can scale well if causal relations are linear (Kalisch and Bühlman, 2007; Ramsey et al., 2017) although their output may not be informative enough (e.g. containing bidirectional edges). In contrast, by making strong assumptions about the functional form of the causal process, FCM can better identify the causal direction (Hyvärinen and Pajunen, 1999; Zhang et al., 2015), although FCM methods usually do not scale well (Glymour et al., 2019). Besides, if the assumed FCM is too restrictive to be able to approximate the true data generating process, the results may be misleading.

There are also various CD methods for timeseries (Chu et al., 2008; Runge et al., 2019b; Runge, 2020; Entner and Hoyer, 2010; Malinsky and Spirtes, 2018, 2019; Hyvärinen et al., 2010; Peters et al., 2013; Pamfil et al., 2020). These methods take in consecutive blocks of observations and output a Full Time Graph (Peters et al., 2017), which contain not only the variables in the system but also their temporally-lagged versions. Although methods for timeseries may be better than cross-sectional ones, they are still not ideal for longitudinal data where sparse observations with potentially missing values come from more than one subject.

Granger Causality: GC (Granger, 1969, 1980) checks for dependence between variables' timeseries, after accounting for other available information. Temporal dependence is thus linked to causation by the "Common Cause Principle": two dependent variables are causally related (one causes the other, or both share a common cause) (Peters et al., 2017). Checking pairwise dependence in GC can be efficient, but often yields false positives because other variables in the system are not accounted for. In contrast, multivariate GC can account for common causes and therefore is more accurate but also more computationally demanding (Eichler, 2007, 2012). In practice, multivariate GC may be infeasible for a large set of variables and more efficient approaches (Basu et al., 2015; Huang and Kleinberg, 2015) were developed to deal with this challenge. Recently, more general GC tests based on neural networks (Tank et al., 2021; Nauta et al., 2019; Khanna and Tan, 2020) have been proposed which outperform vector auto regressive (VAR) linear GC (Glymour et al., 2019). Scaling these neural-network based GC methods to handle a large number of variables is still a concern.

Missing data: For cross-sectional studies, missing values can be imputed, which may result in data contradicting the causal processes if imputation is done naively. Alternatively, observations with missing values can be removed (list-wise deletion), which can lead to the omission of vast amounts of valuable datapoints. Test-wise Deletion PC (TDPC) (Strobl et al., 2018) is more data-efficient than list-wise deletion but may produce spurious edges when missingness is not completely at random (Tu et al., 2019). Missing-Value PC (MVPC) (Tu et al., 2019) corrects TDPC's output to account for different missingness scenarios. Although, tackling missingness in data through imputation has been studied extensively (Ma et al., 2019, 2020; Ma and Zhang, 2021; Morales-Alvarez et al., 2022) in the context of cross-sectional studies, to the best of our knowledge, no existing method addresses missingness in longitudinal studies for the CD task.

3. Method

Both cross-sectional CD methods (multiple subject, single timepoint data) and timeseries CD methods (single subject, multiple timepoints data) are ill-suited for longitudinal studies (multiple subjects, multiple timepoints data). Besides, prior methods often assume timeseries are infinitely long (i.e. unlimited history), regularly sampled, and without missing values. Thus, they may not work for real-world datasets when observation history per subject is limited, irregular, and riddled with missing values. The next few sections show (1) how GLACIAL handles longitudinal data, (2) how GLACIAL deals with irregularly sampled timepoints containing missing values, and (3) GLACIAL's post-processing strategies to account for limited history of observed timeseries.

Causal discovery is impossible without assumptions. GLACIAL assumes causal faithfulness, no hidden confounder, acyclicity (DAG, hence no feedback effect) and no instantaneous effects (the first three assumptions are standard in CD literature, c.f. (Pearl, 2009)). GLACIAL does not assume stationarity unlike linear GC.

3.1 Longitudinal Study Set-up

In a longitudinal imaging study, there are *multiple* subjects who are sparsely scanned during a limited number of visits. Let $\mathbf{X}_t$ and $\mathbf{Y}_t$ be time-varying variables (e.g., two image-derived biomarkers, or an image-derived biomarker and a clinical score) indexed with non-negative integer $t \in \{0, \dots T-1\} = [T]$. We use super-script notation to indicate history: $\mathbf{X}^t = \{\mathbf{X}_0, \dots, \mathbf{X}_{t-1}\}$. $\mathbf{\Omega}^t = \mathbf{X}^t \cup \mathbf{Y}^t \cup \dots$ is the union of histories of all variables. The data from subject i with T_i observations ($T_i \leq T$) is $\mathbf{\Omega}^{T_i}$. The whole longitudinal dataset is $\{\mathbf{\Omega}^{T_i};\ \ i \in 1, \dots, N\}$. The number of observations, T_i, is usually less than 10 (sparse) and can be as low as 1 or 2. The number of subjects, N, is often less than 10,000. The $\mathbf{\Omega}^{T_i}$ matrices may contain missing values.

3.2 Granger Causality Formulation

A popular GC test is based on comparing the mean squared error (MSE) achieved by two predictors (Granger, 1980). In the GC MSE formulation, we conclude that "Y causes X" if:

$$\begin{aligned}\delta_t(X|Y) = \mathsf{MSE}(\mathbf{X}_t, \mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t \setminus \mathbf{Y}^t]) \\ - \mathsf{MSE}(\mathbf{X}_t, \mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t]) > 0, \forall t \in [T]\end{aligned} \tag{1}$$

where $\mathbb{E}$ denotes (conditional) expectations. Eq 1 simply calculates the MSE difference between two optimal (in an MSE sense) predictors of X (see Appendix A.1 and (Granger, 1980)). The first predictor (i.e. $\mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t \setminus \mathbf{Y}^t]$) is not given information about Y. The second predictor (i.e. $\mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t]$) is given all past information, including about Y. Since $\delta_t(X|Y) > 0, \forall t$, Eq 1 can be adapted for longitudinal data as:

$$\Delta\mathsf{MSE}(X|Y) = \mathbb{E}_i\Big[\frac{1}{T_i}\sum_{t=0}^{T_i-1}\delta_t(X|Y)\Big] > 0 \tag{2}$$

Relying on the assumption that statistical dependence implies a causal link (Reichenbach, 1956), when past values of Y predict future values of X (dependence): (1) X causes Y

OR (2) Y causes X OR (3) X and Y have a common cause. With sufficiently high sampling rate, causes are observed to occur before effects in time, thus ruling out (1). The no hidden confounder assumption rules out (3). Hence, a positive test implies that "Y causes X". The next section details how this test can be done in practice when the optimal predictors are not given. We are particularly interested in the setting with multiple observed independent subject trajectories.

3.3 Choice of Predictor

We can approximate the MSE-optimal predictors with neural networks F and G.

$$\begin{aligned}\delta_t(X|Y;F,G) = \mathsf{MSE}(\mathbf{X}_t, F(\mathbf{X}_t;\mathbf{\Omega}^t \setminus \mathbf{Y}^t)) \\ - \mathsf{MSE}(\mathbf{X}_t, G(\mathbf{X}_t;\mathbf{\Omega}^t))\end{aligned} \tag{3}$$

$$\Delta\mathsf{MSE}(X|Y;F,G) = \mathbb{E}_i\Big[\frac{1}{T_i}\sum_{t=1}^{T_i}\delta_t(X|Y;F,G)\Big] \tag{4}$$

To calculate $\delta_t(X|Y;F,G)$, we first have to train the forecasting neural networks. Once trained, the neural networks can be used to calculate $\Delta\mathsf{MSE}(X|Y;F,G)$ on hold-out test subjects. Thus, the predictors' performance depends on the training data, optimization, network initialization, and other implementation details. Even with the best optimizer and initialization procedure, a bad training-test split could, for instance, result in a sub-optimal model and consequently false causal link estimates. For more robust causal discovery, in GLACIAL, we repeat the estimation of $\Delta\mathsf{MSE}(X|Y;F,G)$ multiple times using different random splits of data and test that $\Delta\mathsf{MSE}$ is positive on average using a statistical test.

We use a *single* forecasting recurrent neural network (RNN) (Graves et al., 2008) in place of all predictors. The RNN is trained to predict the next step values of all the variables, $\mathbf{\Omega}_t$, given all available past values, $\mathbf{\Omega}^t$. We adopt the RNN model from (Nguyen et al., 2020) since it implements model interpolation to handle missing values. In particular, if there are missing values at time t, they can be replaced by the RNN prediction, $\widehat{\mathbf{\Omega}}_t$ (model interpolation). This timepoint with interpolated values is then concatenated with $\mathbf{\Omega}^t$ to form $\mathbf{\Omega}^{t+1}$ which is subsequently used to predict values at time t+1. In this way, the RNN is not aware whether the features are observed/measured or imputed. Missing values are ignored when calculating training loss and estimating $\delta_t(X|Y;F,G)$ on hold-out subjects (Eq 3). Training follows (Nguyen et al., 2020) so that even data containing missing values can be used for RNN training. Note that the choice of the neural network is not very critical. Given that the network can fit the data well, any neural network model that forecasts future values from past values and implements model interpolation should work in GLACIAL.

Input Feature Dropout: Training separate neural networks to compute $\Delta\mathsf{MSE}(X|Y;F,G)$ for each variable pair would make this approach infeasible for medical imaging studies with numerous variables. This is because the number of networks required would be proportional to the number of variables squared. Instead, we propose to train a single multi-task (i.e., multi-output) RNN, $F(\cdot;\theta)$, to approximate $\mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t \setminus \mathbf{Y}^t]$ and $\mathbb{E}[\mathbf{X}_t|\mathbf{\Omega}^t]$, for all predicted variables $\mathbf{X}_t$. A similar technique has been shown to allow a single model to learn multiple predictive functions (Nguyen et al., 2024). The RNN acts as the former when Y is masked out of the input vector and acts as the latter when the input is complete. To obtain a model

that can produce accurate predictions under these scenarios, during training, we augment each mini-batch by dropping out subject variables from the input features.

Implementation Details: The same settings of GLACIAL are used in all experiments. We used repeated 5-fold cross-validation to split a dataset into training, validation, and test sets with a 3:1:1 ratio. The RNN is trained to minimize next-step prediction error using Adam (Kingma and Ba, 2014), L2 loss, and a learning rate of 3E-4. The RNN has one hidden layer of size 256. Training was done on an NVIDIA TITAN Xp GPU. The validation set is used for early stopping. Cross-validation is repeated 4 times, resulting in 20 different splits of data. We find 4 repetitions to strike a good balance between robustness and speed. Running more repetitions might slightly improve the results when missingness is severe but at a higher computational cost (see Appendix C.1). We perform a t-test on the $\Delta\mathsf{MSE}$ statistic and use the significance level threshold of 0.05.

3.4 Post-Processing

GC assumes history of the timeseries is infinite. When observations are finite as in real-world longitudinal studies, GC may draw wrong conclusions. E.g., consider following deterministic system:

$$\begin{aligned}\mathbf{Y}_t &= a\mathbf{Y}_{t-1} + b\mathbf{Y}_{t-2}\\ \mathbf{X}_t &= c\mathbf{Y}_{t-2}.\end{aligned}$$

In this system, Y causes X since manipulating Y will change the value of X. By the same logic, X is not the cause of Y because manipulating X will not change Y.

When history is infinite, GC works as expected

$$\begin{aligned}&\mathbb{E}[\mathbf{Y}_t|\mathbf{X}^t,\mathbf{Y}^t] = \mathbb{E}[\mathbf{Y}_t|\mathbf{Y}^t] = \mathbf{Y}_t\\ &\mathsf{MSE}(\mathbf{Y}_t,\mathbb{E}[\mathbf{Y}_t|\mathbf{Y}^t]) = \mathsf{MSE}(\mathbf{Y}_t,\mathbb{E}[\mathbf{Y}_t|\mathbf{X}^t,\mathbf{Y}^t]) = 0\\ &\Rightarrow X \text{ does not cause } Y \quad \text{(correct)}\end{aligned}$$

However, when only 1 past timepoint is given (finite history), GC draws a wrong inference.

$$\begin{aligned}&\mathsf{MSE}(\mathbf{Y}_t,\mathbb{E}[\mathbf{Y}_t|\mathbf{Y}_{t-1}]) > \mathsf{MSE}(\mathbf{Y}_t,\mathbb{E}[\mathbf{Y}_t|c\mathbf{Y}_{t-2},\mathbf{Y}_{t-1}]) = \mathsf{MSE}(\mathbf{Y}_t,\mathbb{E}[\mathbf{Y}_t|\mathbf{X}_{t-1},\mathbf{Y}_{t-1}])\\ &\Rightarrow X \text{ causes } Y \quad \text{(incorrect)}\end{aligned}$$

Thus, GC may detect edges in both direction ($X \to Y$ and $Y \to X$) for a pair of variables when limited history is given. It can be shown in a similar fashion that if X causes Y and Y causes Z (X is the indirect cause of Z), Y will not be able to shield Z from X if only limited history is given. Thus, GC will also detect edges for indirect causes in both direction ($X \to Z$ and $Z \to X$).

GLACIAL includes two additional post-processing steps to remove these false positives. Let $S(X|Y)$ be the statistic (e.g. the t-test) that tests for the positivity of $\Delta\mathsf{MSE}$ from several train/test splits. Thus $S(X|Y)$ can be viewed as a test for whether Y causes X.

1. Orient bidirectional edge: If $S(X|Y) < S(Y|X)$ remove $Y \to X$, else remove $X \to Y$. This step is similar to prior work such as (Hoyer et al., 2008; Zhang and Hyvärinen, 2009a; Janzing et al., 2012; Kocaoglu et al., 2017) which leverages causal asymmetry to

Algorithm 1 GLACIAL

In: Data splits $(D_1^{train}, D_1^{test}), \ldots, (D_n^{train}, D_n^{test})$
Out: Causal graph G
Step 1: Association check using the GC MSE test
For each data split D_i
  Fit RNN model F_i using D_i^{train}
  For each variable pair (u, v)
    Calculate ΔMSE[u, v, i] using F_i and D_i^{test};
For each variable pair (u, v)
  t-statistic, p-value = t-test(ΔMSE[u, v, $*$]); (t-test across data splits)
  If p-value < threshold
    Add $u \to v$ to G; S[u, v] = t-statistic;
Step 2: Orient bidirectional edges
For each bidirectional pair $u \to v$ and $v \to u$ in G
  If $S[u, v] < S[v, u]$
    Remove $u \to v$ from G; ($v \to u$ has stronger effect)
  Else
    Remove $v \to u$ from G; ($u \to v$ has stronger effect)
Step 3: Prune indirect causes
For each $u \to v$ in G
  For each path $p = (u{=}w_0, w_1, \ldots, w_k{=}v)$
    If $S[u, v] < S[w_j, w_{j+1}] \ \ \forall j \in \{0, \ldots, k{-}1\}$
      Remove $u \to v$; break;
  For each path $p = (v{=}w_0, w_1, \ldots, w_k{=}u)$
    If $S[u, v] < S[w_j, w_{j+1}] \ \ \forall j \in \{0, \ldots, k{-}1\}$
      Remove $u \to v$; break;

determine the causal direction (the direction with the bigger effect is regarded as the causal direction). T-statistic has been shown to be informative for causal discovery (Weichwald et al., 2020). Appendix A.2 presents a mathematical justification for this heuristic.

2. Remove indirect edge: Remove edge $X \to Y$ if there exists an alternative path $(X{:=}U_0, U_1, \ldots, Y{:=}U_k)$ from X to Y or a path $(Y{:=}U_0, U_1, \ldots, X{:=}U_k)$ from Y to X if

$$S(Y|X) < \min(\{S(U_{j+1}|U_j); \ \ j \in 0, \ldots, k{-}1\}).$$

Intuitively, if there is an alternative path on which the effect of the weakest edge is greater than the effect of $X \to Y$ then X is likely an indirect cause of Y. A complete description of GLACIAL is shown in Algorithm 1.

3.5 Runtime Complexity

Since GLACIAL uses a single multi-task RNN to check relationships between all variable pairs, the number of RNNs trained by GLACIAL is independent of the number of variables and is equal to the number of data splits (see Algorithm 1 STEP 1). For example, with 4 repetitions of 5-fold cross-validation, GLACIAL needs to train 20 different RNNs. This

number is the same whether there are 10 or 100 variables. Obviously, having more variables will lead to longer execution time per batch but much of the computation is parallelizable (as long as the batch fits into GPU memory). Therefore, the runtime complexity is mostly dominated by the number of RNNs that must be trained.

4. Experiments

4.1 Experimental Set-up

In addition to the problems listed in Section 3, CD methods often struggle when (1) relationships are non-linear, (2) the number of variables is large, or (3) a node has many parents. The subsequent experiments are designed to show GLACIAL's efficacy and to show that GLACIAL is less affected by these problems. First, the simulations include both non-linear trajectories and linear random-walk trajectories. GLACIAL is also applied on real multivariate medical imaging data which most likely include non-linear trajectories. Second, there is a simulation with a moderate-size graph consisting of 39 nodes to demonstrate scalability. Third, the simulation with the 39-node graph includes one node (i.e. 22) with 18 direct causes.

4.1.1 Baselines

We benchmark GLACIAL against both CD methods for cross-sectional data and CD methods for timeseries. Only representative and competitive baselines are shown (see Appendix C.2 for the remaining baselines).

CD Methods for Cross-Sectional Data: We compare against PC, FCI (Spirtes et al., 2000), GFCI (Ogarrio et al., 2016), and *Sort-N-Regress* (Reisach et al., 2021) (SnR). GFCI combines GES and FCI into a single algorithm. SnR is a simple baseline to ensure that benchmarked approaches go beyond exploiting differences in variables' marginal variance (Reisach et al., 2021). As these approaches assume independent observations, only first timepoints (observations) of subjects are used. In most longitudinal studies, subjects are guaranteed to have first timepoints (but not other timepoints). Hence, using the first timepoints will result in the most number of independent timepoints with the least amount of missing data in real-world datasets. Besides, using all timepoints led to worse performance in our preliminary experiments using simulated data. Similar to (Shen et al., 2020), GFCI is run multiple times (i.e. 20) using different bootstraps of subjects' first timepoints, resulting in multiple graphs. Only edges appearing in more than half of the resultant graphs are kept in final graph. Using a higher threshold (80%) led to worse result (see Appendix C.4).

CD Methods for Timeseries Data: We also adopt SVAR-GFCI (Malinsky and Spirtes, 2019), PCMCI+ (Runge, 2020), DYNO-TEARS (Pamfil et al., 2020), and several GC-based approaches as baselines. GC-based approaches include linear GC and more recent neural GC tests: cMLP and cLSTM from (Tank et al., 2021), TCDF from (Nauta et al., 2019), SRU and eSRU from (Khanna and Tan, 2020). For linear GC, F-statistic was used to test for presence of edges using the same threshold as in GLACIAL. For longitudinal data, one could either (1) estimate one causal graph for each subject and aggregate the graphs or (2) estimate just one graph using concatenated data from all the subjects. Since (1) often fails when the number of timepoints per subject is sparse, (2) was used instead.

Causal discovery using concatenated subjects' data has been investigated in (Di et al., 2019; Qing et al., 2021). Besides, linear GC could output false positives when timeseries are non-stationary (He and Maekawa, 2001). One could make the timeseries stationary by calculating the difference (Evgenidis et al., 2017) or the log difference between timepoints (Stock and Watson, 2012). However, using differences led to worse results so we report the results using the original timeseries instead.

The input to SVAR-GFCI and PCMCI+ are also the concatenated timeseries from all the subjects. For DYNOTEARS which can accept timeseries from multiple subjects, the timeseries are not concatenated. The hyper-parameters of SVAR-GFCI, PCMCI+, DYNOTEARS and neural GC tests are selected based on the suggestions in their original publications.

Missing data: For data with missingness, TDPC (Strobl et al., 2018) and MVPC (Tu et al., 2019) are used instead of GFCI. For a dataset, each algorithm is run 20 times and the results are aggregated using the same 50% threshold. As far as we know, there is no prior work on applying causal discovery methods to timeseries data with missing values. Therefore, we used linear interpolation to fill out missing values in the data before applying these methods (linear/neural GC, SVAR-GFCI, PCMCI+, and DYNOTEARS). It may not be feasible to apply more complex interpolation methods since the number of timepoints of a subject can be as low as 2 (after discounting missing values).

4.1.2 Simulated Data

The sample size in the simulations was set to 2000 subjects, roughly the size of the ADNI dataset. Only six timepoints are extracted from each subject's timeseries to simulate sparse observations (see Appendix C.5 for results with 24 timepoints). We consider two scenarios. First, the temporal dynamics are parameterized via the sigmoid function, which is a widely used model for the trajectories of biomarkers, e.g., in Alzheimer's disease (Jack Jr et al., 2013). In the second scenario, we implement random-walk series. We experimented with 3 different structural causal models (SCMs), one linear SCM and two non-linear SCMs. See Appendix B for further details. Results for the non-linear SCMs are shown in Appendix C.3. As causal structure of simulated data may leak through variables' marginal variance, the data are standardized to zero-mean and unit-variance to prevent CD algorithms from gaming the simulated data (Reisach et al., 2021).

Fig 2 shows the causal graphs used for generating the synthetic data. The first graph (7 nodes) contains all the basic structures, namely chain, fork, and collider. The second graph (39 nodes) is used to demonstrate GLACIAL's scalability. This graph is inspired by the RTK/RAS signaling pathway in oncology and is taken from (Sanchez-Vega et al., 2018). The second graph is a realistic target that a causal discovery algorithm should be able to find from observational data. Since the shape of the evolution of signaling proteins is not known, we use Gaussian random-walk as the sample path function. To simulate missingness (completely at random; MCAR), the values for each timeseries of a subject are independently dropped at fixed rate $p \in \{0.1, 0.3, 0.5\}$. Since real data missingness may be more adverse than MCAR, results on simulated missing data are optimistic estimates of performance. The missingness rate is chosen to match the rate in real data. Since

values from different timeseries are dropped independently, the resulted data could contain subjects with all timepoints having at least one missing values.

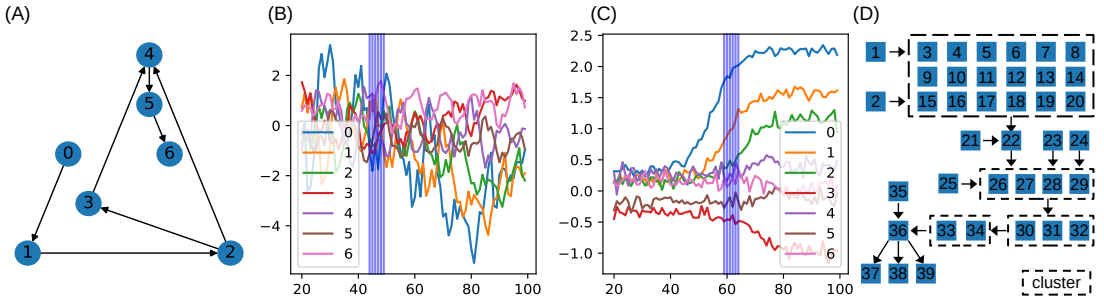


Figure 2: Simulation. (A) 7-node graph having all basic structures (chain, fork, collider). (B) Subject with random-walk trajectories and linear SCM (data before standardizing to zero mean and unit variance). Only timepoints under vertical lines are observed. (C) Subject with sigmoid trajectories and linear SCM. (D) More realistic 39-node graph resembling the RTK/RAS signaling pathway. Nodes in the same cluster have the same causal relations.

4.1.3 Real-world Data from an Alzheimer's Disease Study

We use ADNI (Jack Jr et al., 2008), a longitudinal study of Alzheimer's disease (AD) that consists of 1789 subjects and includes multi-modalities of medical images. Each subject in ADNI has about 7 timepoints on average. The ADNI study tracks multiple AD biomarkers such as region-of-interest (ROI) volumes (e.g. hippocampal) derived from structural MRI scans, cognitive tests (e.g. ADAS13), proteins (e.g. amyloid beta) derived from cerebral spinal fluid samples, and molecular imaging that captures the brain's metabolism (e.g. FDG PET). The missingness rates vary for different biomarkers, ranging from 30% (ADAS13) to around 80% (FDG PET). The variables are shown in Fig 7c and described in Appendix D. For ease of interpretability, we only experiment with summary statistics (volumetric measurements) of medical images. However, GLACIAL can be easily extended to find causal relations between more fine-grain variables in medical images.

4.1.4 Metrics

F1-score, which is the harmonic mean of precision and recall, is used to quantify different approaches' performance. Note that we assume that there is a ground-truth (directed) graph that describes causal relations. Each method will also return a list of directed edges between variables. Precision is the ratio of correctly identified edges over all predicted edges, while recall is the ratio of correct edges over all ground-truth edges. A predicted edge is considered incorrect if the edge does not exist in the ground-truth graph or the predicted direction contradicts the ground-truth direction. Thus, a predicted bidirectional edge would be incorrect if the ground-truth edge has only one direction.

5. Results

5.1 Simulated Data

7-node graph: For random-walk data, GLACIAL outperforms the baselines for various lag-times and measurement noise levels (Fig 3, 1st column). Similarly, GLACIAL also outperforms the baselines, for the sigmoid data (2nd and 3rd column). GLACIAL's performance dips (3rd column) when input history (5 years) is shorter than the lag-time (6 or 7 years). This dip is more pronounced when measurement noise is high (3rd column, bottom).

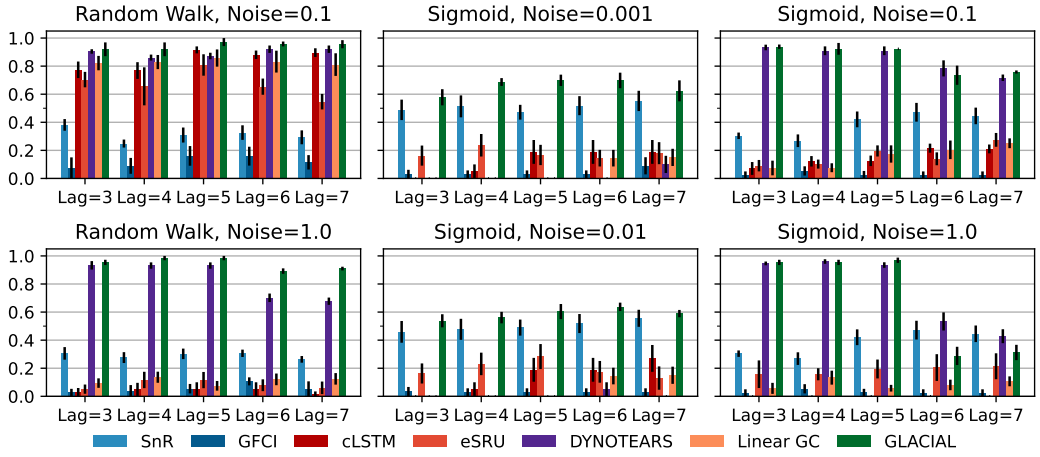


Figure 3: Average F1-scores at different settings of sample path, lag-time and measurement noise (7-node graph). GLACIAL outperforms baselines in most settings (see Appendix C.2 for more comparisons).

DYNOTEARS fails to detect causal relations in systems with almost deterministic dynamics (2nd column) even though it is the best baseline. System with deterministic dynamics is also challenging for linear GC (Peters et al., 2017) although it is slightly better than DYNOTEARS (F1-score < 0.2). Interestingly, GLACIAL still works in these systems (F1-score $= 0.6$). Only GLACIAL manages to consistently beat the strong SnR (*Sort-N-Regress*) baseline.

39-node graph: Although DYNOTEARS is the best baseline for the 7-node graph, its performance on the big graph is worse than linear Granger (Fig. 4). GLACIAL consistently outperforms all baselines on this large graph when the sample path is Gaussian random-walk. GLACIAL performs quite well despite the presence of a cluster of direct causes whose contribution to node "22" may be too small to detect.

Missing data: Fig 5 shows F1-scores at different degrees of missingness. GLACIAL outperforms TDPC and MVPC, CD approaches tailored for missing data, by better exploiting the temporal dynamics within subjects' timeseries. GLACIAL also outperforms CD methods for timeseries such as cLSTM and DYNOTEARS. Although being the best baseline, DYNOTEARS often fails when the missingness level is high (¿0.3). When half of the values

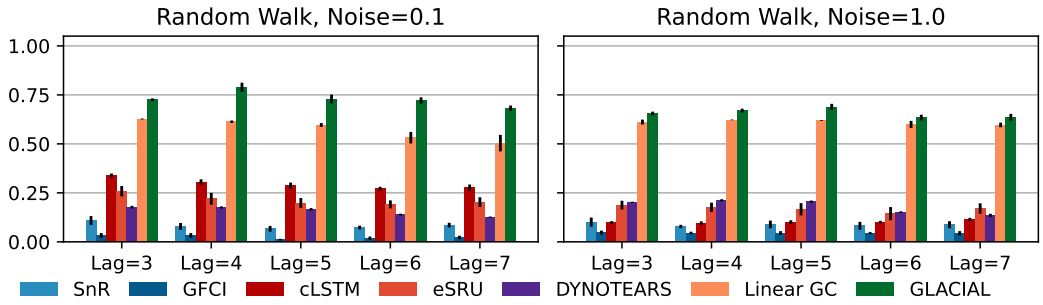


Figure 4: Average F1-scores at different settings of lag-time and measurement noise (39-node graph, Gaussian random-walk). GLACIAL outperforms baselines in most settings (see Appendix C.2 for more comparisons).

are missing (=0.5), GLACIAL can still infer some causal relations. As an aside, GLACIAL's performance on missing data can be improved with more repetitions (see Appendix C.1).

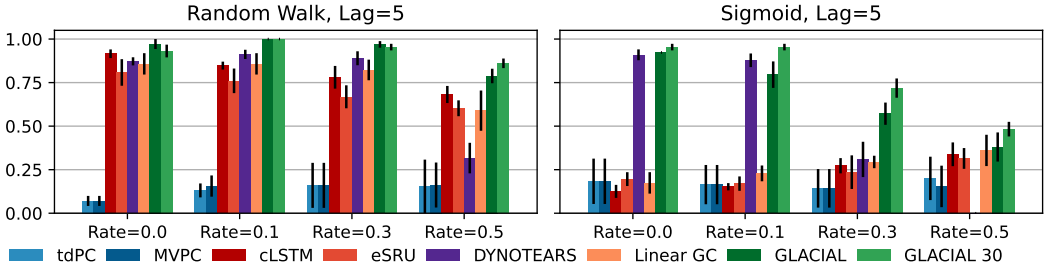


Figure 5: Average F1-scores at various levels of missing at random. Lag-time=5. Noise level = 0.1. GLACIAL usually outperforms baselines. Running GLACIAL for more repetitions (i.e. 30 instead of 4, denoted as GLACIAL 30; see Section 3.3) can improve performance when dealing with missing data.

5.2 GLACIAL's Post-processing Ablation

GLACIAL's first step tests for edges in the causal graph by comparing the difference in MSE on hold-out subjects. However, when test subjects are only sparsely observed for a limited number of times, this step may find spurious edges (edges from effect to cause or edges between indirect cause, e.g. a grand-parent, and effect). To address this problem, GLACIAL has two additional heuristics: one (Step 2) to remove edges from effect to cause and another (Step 3) to prune edges between indirect cause and effect. Fig 6 shows the contribution of these two post-processing heuristics to F1-scores at various lag-times and noise levels (7-node graph simulation). The first heuristic (Step 2) consistently leads to better results. While the second heuristic (Step 3) is beneficial most of the time, it can

sometime result in performance degradation. Thus, when applying GLACIAL to real data, it is recommended to compare the outputs with and without the second heuristic to decide which output is more plausible.

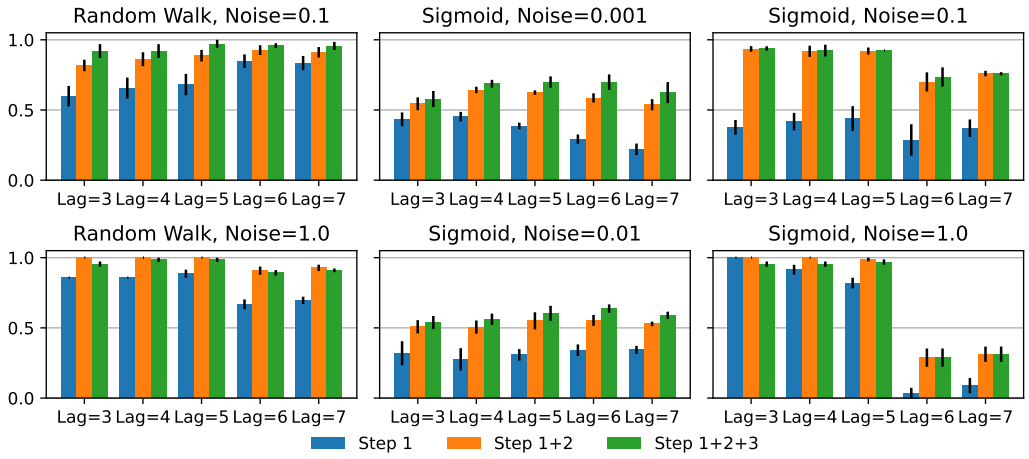


Figure 6: Contribution from GLACIAL's heuristics to F1-scores. 7-node graph simulation.

5.3 GLACIAL's Hyper-parameter Sensitivity Ablation

Since GLACIAL uses neural network for inference, one may think that its results are sensitive to the choice of hyper-parameters. We analyzed GLACIAL's performance as the hyper-parameters vary. Table 1 shows the performance of GLACIAL while varying (1) the number of hidden layers, (2) the size of the hidden layer(s), and (3) the learning rate used. GLACIAL's results seem quite robust to the choice of hyper-parameters.

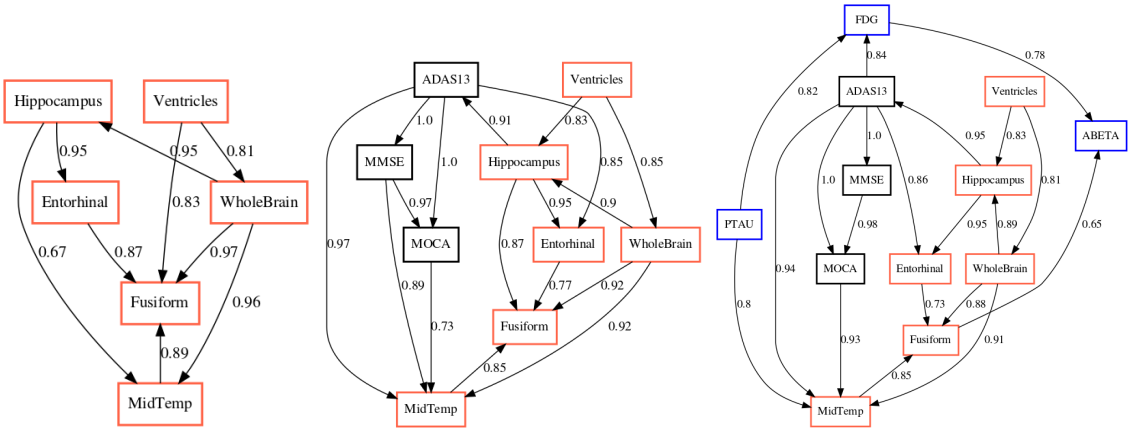
Table 1: GLACIAL's results vary little with different hyper-parameters. Lag-time=5. Noise level = 0.1. L: number of layers, D: size of hidden layer, R: learning rate (D1: 128, D2: 256, D3: 512, R1: 1E-3, R2: 3E-4, R3: 1E-4)

Simulation	**L1-D2-R2**	L1-D1-R2	L1-D3-R2	L2-D2-R2	L1-D2-R1	L1-D2-R3
Random-walk	0.97±0.06	0.96±0.06	0.96±0.06	0.96±0.06	0.92±0.09	0.97±0.06
Sigmoid	0.92±0.00	0.92±0.09	0.91±0.08	0.94±0.03	0.92±0.00	0.92±0.00

5.4 Results on ADNI Data

The output of applying GLACIAL to different sets of ADNI biomarkers are shown in Fig 7. Edge weights denote the frequencies at which edges were detected in multiple runs. Most of the edges are consistently detected across different runs with the exception of "Hippocampus → MidTemp" (67%, Fig 7a) and "Fusiform → ABETA" (65%, Fig 7c). Although

GLACIAL's neural forecasting model assumes MCAR and missingness in ADNI data may be more adverse than that, GLACIAL's result seems promising. There is a high degree of agreement between the 3 graphs which all show the "Ventricle" is a root in the causal graph and "Fusiform" is at the end of the chain, which is aligned with prior work Nestor et al. (2008). The presence of the edge "Hippocampus → Entorhinal" is also consistent with literature Krumm et al. (2016); Planche et al. (2022). In comparison, baselines' outputs are less interpretable (Fig 8; more results are in Appendix D). The outputs of DYNOTEARS and linear Granger contain hardly any edge between ROI volumes while the outputs of cLSTM and eSRU have bidirectional edges.



(a) Only ROI volumes (b) ROI volumes & cognitive tests (c) Extended set of variables

Figure 7: GLACIAL's predicted interaction of ADNI biomarkers. ROI volumes are in red, cognitive tests are in black, and the rest are in blue. ABETA: amyloid beta, PTAU: phosphorylated tau. Edge weights are frequencies at which edges were detected in multiple runs.

One potential issue with GLACIAL's output is that some cognitive scores are causal parents to some ROI volumes. This might be due to the no hidden confounder assumption being violated. Another reason might be measurement noise. For example, based on our understanding of Alzheimer's dynamics, changes in volumetric measurements should cause cognitive decline, and therefore atrophy in MRI biomarkers should precede changes in clinical scores. However, initial volumetric changes may be too small to be captured in MRI images and this may affect the inferred graph.

6. Discussion

Longitudinal studies, in which multiple subjects are sparsely observed for a limited number of times, are particularly common in population health applications. Longitudinal studies often track many variables, which are likely governed by nonlinear dynamics that might have subject-specific idiosyncrasies. Yet, longitudinal studies are not amenable to the popular Granger causality (GC) analysis, since GC was developed to analyze a single mul-

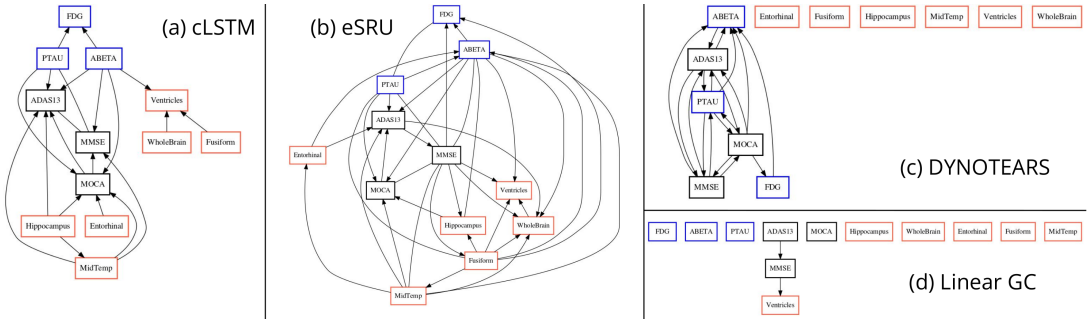


Figure 8: Baseline approaches' predicted interaction of ADNI biomarkers.

tivariate densely sampled timeseries. Furthermore, real-world longitudinal data often suffer from widespread missingness. We developed GLACIAL which combines the GC framework with a machine learning based prediction model to address the need for a method to find causal relations of numerous variables in longitudinal multi-modal medical imaging studies. GLACIAL treats subjects as independent samples and uses average prediction accuracy on hold-out subjects to test for causal relations.

GLACIAL exploits a single multi-task neural network trained with input feature dropout to efficiently probe links. GLACIAL places no restriction on the design of the neural network predictor. This flexibility allows future extensions of our work. For example, Transformers (Vaswani et al., 2017) or Neural ODEs (Chen et al., 2018) can be used instead of RNN.

Although we showed GLACIAL working well in many settings (varying lag-times, noise levels, and missingness degree), there are some questions remaining that need further investigation. Even though the ADNI dataset is small by machine learning standard, many scientific datasets are even smaller so characterizing the performance on small datasets would be interesting. In addition, it would be interesting to use GLACIAL to analyze more medical imaging datasets (Gutiérrez-Zúñiga et al., 2022; Basaia et al., 2022). We focused on continuous variables since they are the most common but extending GLACIAL to discrete variables by adopting techniques in (Peters et al., 2010; Cai et al., 2018; Huang et al., 2018) would make GLACIAL analysis applicable to more longitudinal studies. Furthermore, studying GLACIAL's behaviors under missingness other than MCAR is important despite GLACIAL outputting plausible graphs on real-world data (ADNI). Missing at random (MAR; missingness probabilities depend on the values of the observed features) and missing not at random (MNAR; missingness depends on the values of the ,unobserved features) are different from MCAR, and MNAR in particular may require different treatments. Besides, GLACIAL assumes that there is no feedback, hidden confounder, or instantaneous effect. Thus, before applying GLACIAL, it is critical to verify whether these assumptions are reasonable to ensure that causal relations inferred by GLACIAL are valid. The third assumption in particular requires that the sampling resolution is high enough to capture transient changes (e.g. impulses) or temporal orderings between causal pairs with short time lags. Although causal discovery when some assumptions are violated has been studied in the past (for example, presence of hidden confounders (Spirtes et al., 2000) or presence of

instantaneous effects (Danks and Plis, 2013)), extending these techniques to longitudinal studies is still an open question. We leave these questions for future work.

Similar to other causal discovery methods, GLACIAL can offer valuable insights into relationships between variables. However, intentional or unintentional misuse may lead to harmful consequences. Causal discovery methods, such as GLACIAL, identify potential causal links based on statistical patterns in the data, so biased data collection may lead to misleading findings. These flawed findings may further lead to misguided medical treatments or clinical decisions that can harm vulnerable populations. Moreover, violations of modeling assumptions and measurement errors can contribute to mistakes in the inferred graphs. Thus, it is important to ensure unbiased data collection as well as to corroborate any identified causal links using further experimentation and follow-up study designs to mitigate potential risks when applying causal discovery.

Acknowledgments

This research was supported by NIH grants R01LM012719, R01AG053949, the NSF NeuroNex grant 1707312, and the NSF CAREER 1748377 grant (MS).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals

Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Silvia Basaia, Federica Agosta, Ibai Diez, Elisenda Bueichekú, Federico d'Oleire Uquillas, Manuel Delgado-Alvarado, César Caballero-Gaudes, MariCruz Rodriguez-Oroz, Tanja Stojkovic, Vladimir S Kostic, et al. Neurogenetic traits outline vulnerability to cortical disruption in parkinson's disease. *NeuroImage: Clinical*, 33:102941, 2022.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 16(1):417–453, 2015.

Ruichu Cai, Jie Qiao, K Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Conference on Neural Information Processing Systems*, pages 2671–2679, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Conference on Neural Information Processing Systems*, pages 6572–6583, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Tianjiao Chu, C Glymour, and Greg Ridgeway. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(5), 2008.

David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *Conference on Neural Information Processing Systems: Workshop on Causality*, 2013.

Xin Di, Marie Wölfer, Mario Amend, Hans Wehrl, Tudor M Ionescu, Bernd J Pichler, Bharat B Biswal, and Alzheimer's Disease Neuroimaging Initiative. Interregional causal influences of brain metabolic activity reveal the spread of aging effects during normal aging. *Human Brain Mapping*, 40(16):4657–4668, 2019.

Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

Michael Eichler. Causal inference in time series analysis. *Causality: Statistical Perspectives and Applications*, pages 327–354, 2012.

Doris Entner and P Hoyer. On causal discovery from time series data using fci. *Probabilistic Graphical Models*, pages 121–128, 2010.

Anastasios Evgenidis, Athanasios Tsagkanos, and Costas Siriopoulos. Towards an asymmetric long run equilibrium between stock market uncertainty and the yield spread. a threshold vector error correction approach. *Research in International Business and Finance*, 39:267–279, 2017.

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

C Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

Clive Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 31(5):855–868, 2008.

Raquel Gutiérrez-Zúñiga, Ibai Diez, Elisenda Bueicheku, Chan-Mi Kim, William Orwig, Victor Montal, Blanca Fuentes, Exuperio Díez-Tejedor, Maria Gutiérrez Fernández, and Jorge Sepulcre. Connectomic-genetic signatures in the cerebral small vessel disease. *Neurobiology of Disease*, 167:105671, 2022.

Zonglu He and Koichi Maekawa. On spurious granger causality. *Economics Letters*, 73(3): 307–313, 2001.

MA Hernan and J Robins. *Causal inference: What if, part III.* Chapman & Hall/CRC, 2020.

P Hoyer, D Janzing, J Mooij, J Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. *Conference on Neural Information Processing Systems*, 21, 2008.

B Huang, K Zhang, Yizhu Lin, B Schölkopf, and C Glymour. Generalized score functions for causal discovery. In *International Conference on Knowledge Discovery and Data Mining*, pages 1551–1560, 2018.

Yuxiao Huang and Samantha Kleinberg. Fast and accurate causal inference from time series data. In *International FLAIRS Conference*, 2015.

A Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

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

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal*

of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 27(4):685–691, 2008.

Clifford R Jack Jr, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013.

D Janzing, J Mooij, K Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and B Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Markus Kalisch and P Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2014.

Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Sabine Krumm, Sasa L Kivisaari, Alphonse Probst, Andreas U Monsch, Julia Reinhardt, Stephan Ulmer, Christoph Stippich, Reto W Kressig, and Kirsten I Taylor. Cortical thinning of parahippocampal subregions in very early alzheimer's disease. *Neurobiology of aging*, 38:188–196, 2016.

Songting Li, Yanyang Xiao, Douglas Zhou, and David Cai. Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information. *Physical Review E*, 97(5):052216, 2018.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. In *NeurIPS*, volume 34, pages 27645–27658, 2021.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *ICML*, pages 4234–4243. PMLR, 2019.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. In *NeurIPS*, volume 33, pages 11237–11247, 2020.

D Malinsky and P Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *International Conference on Knowledge Discovery and Data Mining: Workshop on causal discovery*, pages 23–47, 2018.

D Malinsky and P Spirtes. Learning the structure of a nonstationary vector autoregression. In *International Conference on Artificial Intelligence and Statistics*, pages 2986–2994, 2019.

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical Review Letters*, 100(14):144103, 2008.

Pablo Morales-Alvarez, Wenbo Gong, Angus Lamb, Simon Woodhead, Simon Peyton Jones, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Simultaneous missing value imputation and structure learning with groups. In *NeurIPS*, volume 35, pages 20011–20024, 2022.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131 (9):2443–2454, 2008.

Paul Newbold. Feedback induced by measurement errors. *International Economic Review*, pages 787–791, 1978.

Minh Nguyen, Tong He, Lijun An, Daniel C Alexander, Jiashi Feng, and BT Thomas Yeo. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020.

Minh Nguyen, Batuhan K. Karaman, Heejong Kim, Alan Q. Wang, Fengbei Liu, and Mert R. Sabuncu. Knockout: A simple way to handle missing inputs. 2024.

Juan Miguel Ogarrio, P Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

J Peters, D Janzing, and B Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *International Conference on Artificial Intelligence and Statistics*, pages 597–604, 2010.

J Peters, D Janzing, and B Schölkopf. Causal inference on time series using restricted structural equation models. *Conference on Neural Information Processing Systems*, 26, 2013.

J Peters, D Janzing, and B Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

V Planche, JV Manjon, B Mansencal, E Lanuza, T Tourdias, G Catheline, and P Coupé. Structural progression of alzheimer's disease over decades: The mri staging scheme. brain communications, 4 (3), article fcac109, 2022.

Zhao Qing, Feng Chen, Jiaming Lu, Pin Lv, Weiping Li, Xue Liang, Maoxue Wang, Zhengge Wang, Xin Zhang, Bing Zhang, et al. Causal structural covariance network revealing atrophy progression in alzheimer's disease continuum. *Human Brain Mapping*, 42(12): 3950–3962, 2021.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and C Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Alard Roebroeck, Elia Formisano, and Rainer Goebel. Mapping directed influence over the brain using granger causality and fmri. *NeuroImage*, 25(1):230–242, 2005.

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

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, C Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553, 2019a.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019b.

Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific Reports*, 10(1):2975, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006.

P Spirtes, C Glymour, R Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

James H Stock and Mark W Watson. Disentangling the channels of the 2007-2009 recession. Technical report, National Bureau of Economic Research, 2012.

Eric V Strobl, Shyam Visweswaran, and P Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International Journal of Data Science and Analytics*, 6(1):47–62, 2018.

Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(8):4267–4279, 2021.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and K Zhang. Causal discovery in the presence of missing data. In *International Conference on Artificial Intelligence and Statistics*, pages 1762–1770, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

Sebastian Weichwald, Martin E Jakobsen, Phillip B Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *Conference on Neural Information Processing Systems: Competition and Demonstration Track*, pages 27–36, 2020.

Alex Eric Yuan and Wenying Shou. Data-driven causal analysis of observational biological time series. *eLife*, 11:e72518, 2022.

David D Zhang, Harry F Lee, Cong Wang, Baosheng Li, Qing Pei, Jane Zhang, and Yulun An. The causality analysis of climate change and large-scale human crisis. *PNAS*, 108 (42):17296–17301, 2011.

K Zhang and Lai-Wan Chan. Extensions of ica for causality discovery in the hong kong stock market. In *Conference on Neural Information Processing Systems*, pages 400–409, 2006.

K Zhang and A Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–585, 2009a.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Conference on Uncertainty in Artificial Intelligence*, pages 647–655, 2009b.

K Zhang, Zhikun Wang, Jiji Zhang, and B Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Conference on Neural Information Processing Systems*, 2018.

Appendix A. Additional Details about GLACIAL

A.1 Granger Causality MSE Test

Let $\mathbf{X}_t$ and $\mathbf{X}^t = \{\mathbf{X}_0, \ldots, \mathbf{X}_{t-1}\}$ denote a time-varying random variables and its history. Let $\mathbf{\Omega}^t = \mathbf{X}^t \cup \mathbf{Y}^t \cup \ldots$ be the union of all historical variables available at time t. The general GC hypothesis is: Y causing $X \Rightarrow$ there exists some event A and $t \in [T]$ such that

$$\Pr(\mathbf{X}_t \in A | \mathbf{\Omega}^t) \neq \Pr(\mathbf{X}_t \in A | \mathbf{\Omega}^t \setminus \mathbf{Y}^t)$$

Equivalently, we have:

$$\Pr(\mathbf{X}_t | \mathbf{\Omega}^t) = \Pr(\mathbf{X}_t | \mathbf{\Omega}^t \setminus \mathbf{Y}^t), \forall t \in [T]$$
$$\Rightarrow Y \text{ does not cause } X$$

In general it is not possible to compare these conditional probabilities based on observed timeseries data. A practical approach is to compare conditional expectations:

$$\mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t] \stackrel{?}{=} \mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t \setminus \mathbf{Y}^t].$$

Note that conditional expectations can be infeasible to compare, so we often need further assumptions. One observation is that the conditional expectation is the optimal estimator that minimizes the mean square error (MSE). This gives rise to a GC test that compares the MSE of least-squares predictors. In this approach, we conclude that "Y causes X" if:

$$\mathsf{MSE}(\mathbf{X}_t, \mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t]) < \mathsf{MSE}(\mathbf{X}_t, \mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t \setminus \mathbf{Y}^t]) \tag{5}$$

Note that, in above equation the right hand side is larger than or equal to the left hand side because, in general, the least-square loss will be smaller with more history.

In practice, the implementation of the GC MSE test often relies on two more assumptions. The first is the stationarity assumption. That is, we suppose $\mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t]$ and $\mathbb{E}[\mathbf{X}_t | \mathbf{\Omega}^t \setminus \mathbf{Y}^t]$ are independent of t. Second, we assume the stochastic processes are Markovian and thus a finite history (often just the prior timepoint) is sufficient for making the least-square forecast. In our framework, the Markovian assumption can be relaxed because the RNN forecast model can digest all available history.

A.2 GLACIAL's First Heuristic Justification

In this section, we will provide some justification for our post-processing heuristic that removes one of the arrows of a bi-directional edge. We will consider a scenario where only one previous timepoint is given to predict the future timepoint. Furthermore, we will suppose that we have infinite data and infinite capacity models that allow us to estimate the unknown parameters and conditional expectations exactly.

We will assume following data generation process (Fig 9) with constants $a, b, c > 0$ and independent and identically-distributed, zero-mean additive errors $\epsilon_{X,t}, \epsilon_{Y,t}$

$$X_t := aX_{t-1} + bX_{t-2} + \epsilon_{X,t} \tag{6}$$
$$Y_t := cX_{t-1} + \epsilon_{Y,t} \tag{7}$$
$$\mathbb{E}[\epsilon_{X,t}^2] = \sigma_X^2 > 0; \quad \mathbb{E}[\epsilon_{Y,t}^2] = \sigma_Y^2 > 0 \tag{8}$$

Back up 1 step in time

$$X_{t-1} = aX_{t-2} + bX_{t-3} + \epsilon_{X,t-1} \tag{9}$$

$$Y_{t-1} = cX_{t-2} + \epsilon_{Y,t-1} \tag{10}$$

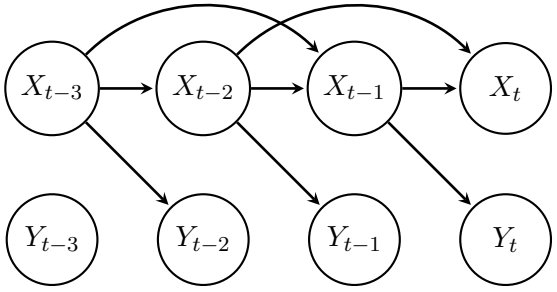


Figure 9: **Data generation.** X is an order-2 auto-regressive process and X causes Y ($X \to Y$).

Computing $\mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}, Y_{t-1}])$

$$X_{t-2} = \frac{1}{c}\Big(Y_{t-1} - \epsilon_{Y,t-1}\Big) \qquad \text{from (10)} \tag{11}$$

$$\begin{aligned} X_t &= aX_{t-1} + \frac{b}{c}\Big(Y_{t-1} - \epsilon_{Y,t-1}\Big) + \epsilon_{X,t} \qquad \text{from (6) and (11)} \\ &= aX_{t-1} + \frac{b}{c}Y_{t-1} - \frac{b}{c}\epsilon_{Y,t-1} + \epsilon_{X,t} \end{aligned} \tag{12}$$

Since $\epsilon_{Y,t-1}$, $\epsilon_{X,t}$ are independent of X_{t-1}, Y_{t-1}, from (12):

$$\Rightarrow \mathbb{E}[X_t|X_{t-1}, Y_{t-1}] = aX_{t-1} + \frac{b}{c}Y_{t-1} \tag{13}$$

$$\Rightarrow X_t - \mathbb{E}[X_t|X_{t-1}, Y_{t-1}] = -\frac{b}{c}\epsilon_{Y,t-1} + \epsilon_{X,t} \tag{14}$$

$$\begin{aligned} \mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}, Y_{t-1}]) &= \mathbb{E}[(X_t - \mathbb{E}[X_t|X_{t-1}, Y_{t-1}])^2] \\ &= \mathbb{E}\Big[\Big(-\frac{b}{c}\epsilon_{Y,t-1} + \epsilon_{X,t}\Big)^2\Big] = \frac{b^2}{c^2}\sigma_Y^2 + \sigma_X^2 \end{aligned} \tag{15}$$

Computing $\mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}])$

$$X_{t-2} = \frac{1}{a}X_{t-1} - \frac{b}{a}X_{t-3} - \frac{1}{a}\epsilon_{X,t-1} \quad \text{from (9)} \tag{16}$$

$$\begin{aligned} X_t &= aX_{t-1} + b\Big(\frac{1}{a}X_{t-1} - \frac{b}{a}X_{t-3} - \frac{1}{a}\epsilon_{X,t-1}\Big) + \epsilon_{X,t} \quad \text{from (6) and (16)} \\ &= \Big(a + \frac{b}{a}\Big)X_{t-1} - \frac{b^2}{a}X_{t-3} - \frac{b}{a}\epsilon_{X,t-1} + \epsilon_{X,t} \end{aligned} \tag{17}$$

Since $\epsilon_{X,t-1}$, $\epsilon_{X,t}$ are independent of X_{t-1}, and assuming the correlation between X_{t-1} and X_{t-3} is weak, from (17):

$$\mathbb{E}[X_t|X_{t-1}] \approx \Big(a + \frac{b}{a}\Big)X_{t-1} \tag{18}$$

$$X_t - \mathbb{E}[X_t|X_{t-1}] \approx -\frac{b^2}{a}X_{t-3} - \frac{b}{a}\epsilon_{X,t-1} + \epsilon_{X,t} \tag{19}$$

$$\begin{aligned}
\mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}]) &\approx \mathbb{E}[(X_t - \mathbb{E}[X_t|X_{t-1}])^2] \\
&\approx \mathbb{E}\Big[\Big(-\frac{b^2}{a}X_{t-3} - \frac{b}{a}\epsilon_{X,t-1} + \epsilon_{X,t}\Big)^2\Big] && (20)\\
&\approx \frac{b^4}{a^2}\mathbb{E}[X_{t-3}^2] + \frac{b^2}{a^2}\sigma_X^2 + \sigma_X^2 \quad \text{(independent)} && (21)
\end{aligned}$$

Estimating $\Delta\mathsf{MSE}(X|Y)$

$$\begin{aligned}
\Delta\mathsf{MSE}(X|Y) &= \mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}]) - \mathsf{MSE}(X_t, \mathbb{E}[X_t|X_{t-1}, Y_{t-1}]) && (22)\\
&\approx \Big(\frac{b^4}{a^2}\mathbb{E}[X_{t-3}^2] + \frac{b^2}{a^2}\sigma_X^2 + \sigma_X^2\Big) - \Big(\frac{b^2}{c^2}\sigma_Y^2 + \sigma_X^2\Big) && (23)\\
&\approx \frac{b^4}{a^2}\mathbb{E}[X_{t-3}^2] + \frac{b^2}{a^2}\sigma_X^2 - \frac{b^2}{c^2}\sigma_Y^2 && (24)
\end{aligned}$$

Computing $\mathsf{MSE}(Y_t, \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}])$

$$\begin{aligned}
&\Rightarrow \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}] = cX_{t-1} \quad \text{from (7)} && (25)\\
&\Rightarrow Y_t - \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}] = \epsilon_{Y,t} \quad \text{from (7)} && (26)\\
&\Rightarrow \mathsf{MSE}(Y_t, \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}]) \\
&= \mathbb{E}[(Y_t - \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}])^2] = \mathbb{E}[(\epsilon_{Y,t})^2] = \sigma_Y^2 && (27)
\end{aligned}$$

Estimating $\mathsf{MSE}(Y_t, \mathbb{E}[Y_t|Y_{t-1}])$

From (9) and (11):

$$X_{t-1} = \frac{a}{c}\Big(Y_{t-1} - \epsilon_{Y,t-1}\Big) + bX_{t-3} + \epsilon_{X,t-1} \tag{28}$$

From (7) and (28):

$$\begin{aligned}
Y_t &= c\Big[\frac{a}{c}\Big(Y_{t-1} - \epsilon_{Y,t-1}\Big) + bX_{t-3} + \epsilon_{X,t-1}\Big] + \epsilon_{Y,t} \\
&= aY_{t-1} - a\epsilon_{Y,t-1} + bcX_{t-3} + c\epsilon_{X,t-1} + \epsilon_{Y,t} && (29)
\end{aligned}$$

Since $\epsilon_{X,t-1}$, $\epsilon_{Y,t}$ are independent of Y_{t-1}, and assuming the correlation between Y_{t-1} and X_{t-3} is weak, from (29):

$$\begin{aligned}
&\Rightarrow \mathbb{E}[Y_t|Y_{t-1}] \approx aY_{t-1} && (30)\\
&\Rightarrow Y_t - \mathbb{E}[Y_t|Y_{t-1}] \approx -a\epsilon_{Y,t-1} + bcX_{t-3} + c\epsilon_{X,t-1} + \epsilon_{Y,t} && (31)
\end{aligned}$$

$$\Rightarrow \mathsf{MSE}(Y_t, \mathbb{E}[Y_t|Y_{t-1}]) = \mathbb{E}[(Y_t - \mathbb{E}[Y_t|Y_{t-1}])^2]$$
$$\approx \mathbb{E}[(-a\epsilon_{Y,t-1} + bcX_{t-3} + c\epsilon_{X,t-1} + \epsilon_{Y,t})^2] \quad (32)$$
$$\approx a^2\sigma_Y^2 + b^2c^2\mathbb{E}[X_{t-3}^2] + c\sigma_X^2 + \sigma_Y^2 \quad \text{(independent)} \quad (33)$$

Estimating $\Delta\mathsf{MSE}(Y|X)$

$$\Delta\mathsf{MSE}(Y|X) = \mathsf{MSE}(Y_t, \mathbb{E}[Y_t|Y_{t-1}]) - \mathsf{MSE}(Y_t, \mathbb{E}[Y_t|X_{t-1}, Y_{t-1}]) \quad (34)$$
$$\approx (a^2\sigma_Y^2 + b^2c^2\mathbb{E}[X_{t-3}^2] + c^2\sigma_X^2 + \sigma_Y^2) - (\sigma_Y^2) \quad (35)$$
$$\approx a^2\sigma_Y^2 + b^2c^2\mathbb{E}[X_{t-3}^2] + c^2\sigma_X^2 \quad (36)$$

Since $\Delta\mathsf{MSE}(Y|X) > 0$, edge $X \to Y$ is always detected. Edge $Y \to X$ is (falsely) detected if $\Delta\mathsf{MSE}(X|Y) > 0$. The heuristic needs to remove this falsely detected edge. We can show that if false detection happens, our heuristic of comparing $\Delta\mathsf{MSE}(Y|X)$ and $\Delta\mathsf{MSE}(X|Y)$ (which is based on the t-statistics) will show the true direction. In other words, let's show that if $\Delta\mathsf{MSE}(X|Y) > 0$, then $\Delta\mathsf{MSE}(Y|X) - \Delta\mathsf{MSE}(X|Y) > 0$.

$$\Delta\mathsf{MSE}(X|Y) > 0 \Leftrightarrow \frac{b^4}{a^2}\mathbb{E}[X_{t-3}^2] + \frac{b^2}{a^2}\sigma_X^2 - \frac{b^2}{c^2}\sigma_Y^2 > 0 \Leftrightarrow b^2\mathbb{E}[X_{t-3}^2] > \frac{a^2}{c^2}\sigma_Y^2 - \sigma_X^2 \quad (37)$$

$$\Delta\mathsf{MSE}(Y|X) - \Delta\mathsf{MSE}(X|Y)$$
$$\approx (a^2\sigma_Y^2 + b^2c^2\mathbb{E}[X_{t-3}^2] + c^2\sigma_X^2) - \Big(\frac{b^4}{a^2}\mathbb{E}[X_{t-3}^2] + \frac{b^2}{a^2}\sigma_X^2 - \frac{b^2}{c^2}\sigma_Y^2\Big) \quad (38)$$
$$\approx a^2\sigma_Y^2 + b^2\mathbb{E}[X_{t-3}^2]\Big(c^2 - \frac{b^2}{a^2}\Big) + c^2\sigma_X^2 - \frac{b^2}{a^2}\sigma_X^2 + \frac{b^2}{c^2}\sigma_Y^2 \quad (39)$$

Let $\mathsf{Diff} = \Delta\mathsf{MSE}(Y|X) - \Delta\mathsf{MSE}(X|Y)$, from (37) and (39)

$$\mathsf{Diff} > a^2\sigma_Y^2 + \Big(\frac{a^2}{c^2}\sigma_Y^2 - \sigma_X^2\Big)\Big(c^2 - \frac{b^2}{a^2}\Big) \qquad + c^2\sigma_X^2 - \frac{b^2}{a^2}\sigma_X^2 + \frac{b^2}{c^2}\sigma_Y^2 \quad (40)$$
$$\Leftrightarrow \mathsf{Diff} > a^2\sigma_Y^2 + a^2\sigma_Y^2 - c^2\sigma_X^2 - \frac{b^2}{c^2}\sigma_Y^2 + \frac{b^2}{a^2}\sigma_X^2 \qquad + c^2\sigma_X^2 - \frac{b^2}{a^2}\sigma_X^2 + \frac{b^2}{c^2}\sigma_Y^2 \quad (41)$$
$$\Leftrightarrow \mathsf{Diff} > 2a^2\sigma_Y^2 > 0 \quad \text{(Q.E.D)} \quad (42)$$

Appendix B. Pseudo-code for Data Generation

Algorithm 2 shows the steps to generate data from a given causal graph G. First, the edge weights are sampled. Then, for each subject, the timeseries are generated in topological order. If a node has no parent, i.e., if it is a source node, its timeseries is specified by the sample path f (Gaussian random-walk or sigmoid). The random-walk function is a conventional choice while the sigmoid function yields trajectories that mimic the evolution of many real-world dynamical systems (Jack Jr et al., 2013). A non-source node's timeseries is the weighted sum of the lagged version of its parents' timeseries. Next, Gaussian measurement noise with standard deviation σ is added to the timeseries. Finally, a discrete set of timepoints within a randomly-chosen observation window are extracted, mimicking a real-world longitudinal study.

- Gaussian random-walk: $f(t) = \sum_{i=0}^{t} \mathcal{N}(0,1)$
- Sigmoid: $f(t) = \frac{A}{1+e^{-k(t-t_0)}}; \quad A \sim \mathsf{Unif}(1,2),\ t_0 \sim \mathsf{Unif}(40,60),\ k \sim \mathsf{Unif}(0.1,0.3)$

For random-walk timeseries, the noise variance σ is either 0.1 or 1.0 (smaller σ has no visible effect). For sigmoid timeseries, $\sigma = 0.001, 0.01, 0.1, 1.0$. Since measurement noise can (1) induce spurious causality between unrelated variables and (2) suppress true causality (Newbold, 1978; Glymour et al., 2019), it is important to benchmark across different levels of measurement noise. For each set of parameters (f, lag-time L, and σ), we generated 5 different randomized datasets so as to estimate the standard error of the performance metrics.

Algorithm 2 Data Generation

In: Causal graph G, sample path f, number of subjects n, number of timepoints m, Lag-time L, measurement noise magnitude σ
Out: Dataset $D = (X_1, \ldots, X_n)$
Sample edge weight
For each $u \to v \in G$
  $s_{uv} \sim \mathsf{Rademacher}()$;
  $m_{uv} \sim \mathsf{Unif}(0.5, 1)$;
  $w_{uv} = s_{uv} * m_{uv}$
Sample observation series of a subject
For each subject i
  For each node v
    $b \sim \mathsf{Unif}(-0.5, 0.5)$ (bias term)
    If v has no parent
      $s_v[t] = b + f[t]$ (time $t \in [0, 100]$)
    Else
      $s_v[t] = b + \sum_{u,(u,v)\in G} w_{uv} * s_u[t - L]$
    $s_v[t] = s_v[t] + \mathcal{N}(0, \sigma)$ (measurement noise)
  $S_i = \{s_u, u \in G\}$ (subject data)
  $t_{\text{start}} \sim \mathsf{Unif}(30, 70)$;
  $t_{\text{end}} = t_{\text{start}} + m$;
  $X_i = S_i[t_{\text{start}} : t_{\text{start}}]$ (extract timepoints within window)

In addition to the linear structural causal model (SCM) shown in Algorithm 2, we also experimented with 2 non-linear SCMs. One SCM takes a polynomial form (Eq. 43) while the other takes a trigonometric form (Eq. 44).

$$c_{uv} \sim \mathsf{Unif}(-1, 1) \qquad p_{uv} \sim \mathsf{Unif}(0.1, 1)$$

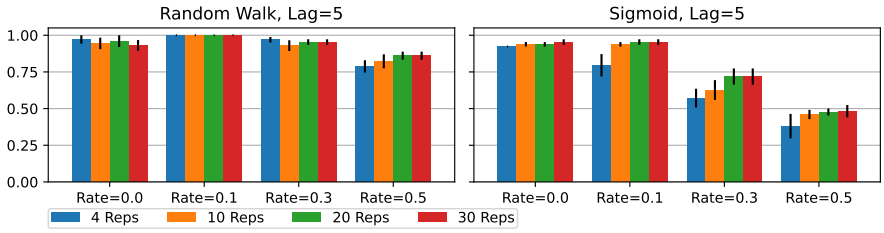
$$s_v[t] = \mathsf{Unif}(-0.5, 0.5) + \mathcal{N}(0, \sigma) + \sum_{u,(u,v)\in G} w_{uv} * |s_u[t - L] + c_{uv}|^{p_{uv}} \tag{43}$$

$$s_v[t] = \mathsf{Unif}(-0.5, 0.5) + \mathcal{N}(0, \sigma) + \sum_{u,(u,v)\in G} w_{uv} * \sin(s_u[t - L] + c_{uv}) \tag{44}$$

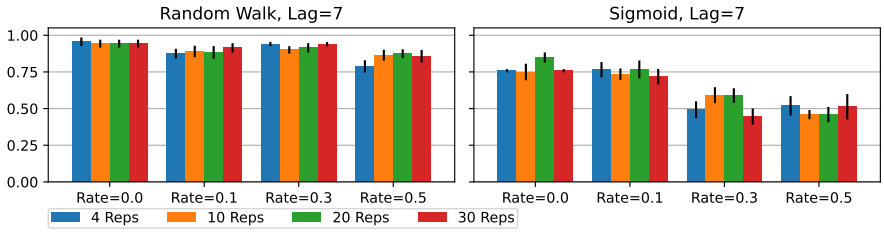
Appendix C. Additional Simulation Results

C.1 More Repetitions

GLACIAL's results in Section 5.1 were obtained by repeating 5-fold cross-validation for 4 times. Increasing the number of repetitions leads to higher F-1 score as shown by the trend in Fig 10. However, the gap between 10 repetitions and 4 repetitions is not large enough to justify the extra computational cost of repeating more than 4 times in Section 5.1. Another interesting trend in Fig 10 is that the gap between 4 repetitions and 30 repetitions is more apparent for missing data. Thus, when the data is noisy (more missing values), more repetitions may yield more accurate results.



(a) Lag-time (L) = 5



(b) Lag-time (L) = 7

Figure 10: More repetitions of cross-validation lead to slightly better result at the expense of running time.

C.2 More Comparisons of GLACIAL against Baselines

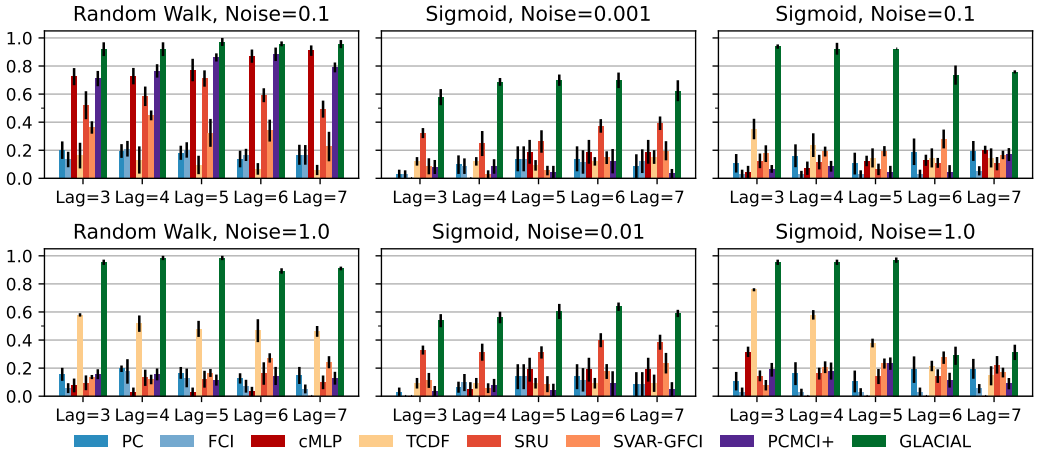


Figure 11: Average F1-scores at different lag-time and measurement noise for 7-node graph. GLACIAL outperforms baselines in most settings of sample path, lag-time, and measurement noise (also see Fig 3).

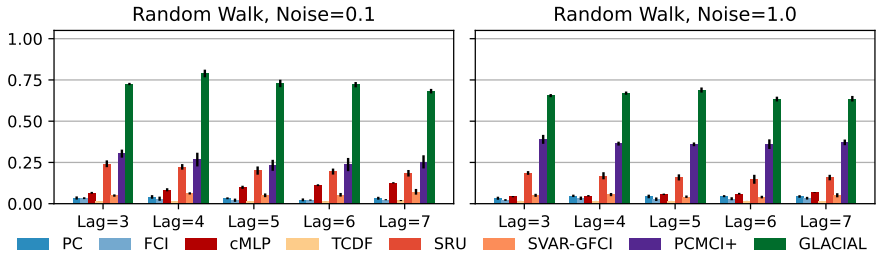


Figure 12: Average F1-scores at different lag-time and measurement noise for 39-node graph (Gaussian random-walk). GLACIAL outperforms baselines in most settings (also see Fig 4).

C.3 Non-linear SCM simulations

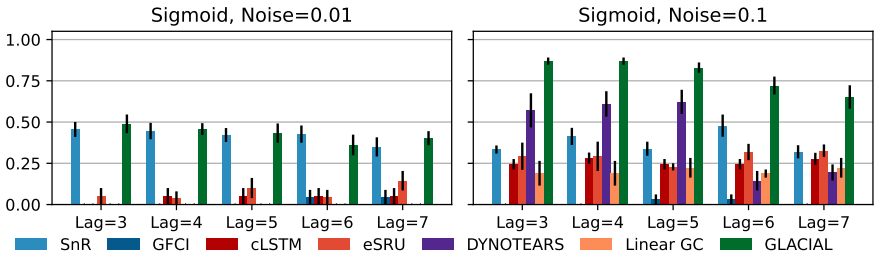


Figure 13: F1-scores at different lag-time and measurement noise (polynomial SCM)

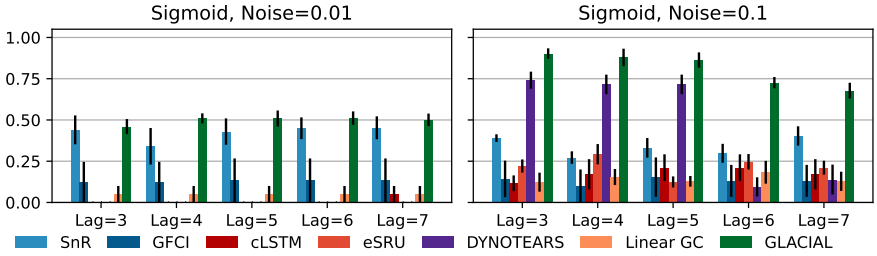


Figure 14: F1-scores at different lag-time and measurement noise (trigonometric SCM)

C.4 Constraint-based Baselines with Higher Threshold

The PC, FCI, and GFCI baselines are run multiple times using different data bootstraps, resulting in multiple graphs. To combine the graphs, we only retain edges that appear more than half of the runs. This procedure is similar to (Shen et al., 2020) although they used a more conservative threshold (0.8) in their work. Fig 15 shows the results of the constraint-based baselines when the threshold of 0.8 (80%) is used. Compared to the results in Fig 3, this more conservative threshold led to worse performance in the baselines.

C.5 More Densely Sampled Data

In Section 5.1, each subject only has 6 timepoints (sparse observations) so linear GC did not work well. Thus, we investigated a scenario more favorable for linear GC where for when subjects have more timepoints (i.e. 24 timepoints). With more timepoints, linear GC results improve slightly but are still worse than that of GLACIAL (Fig 16). Additionally, using GC to estimate one causal graph for each subject could not find the correct graph even with 24 timepoints (hence, result not reported).

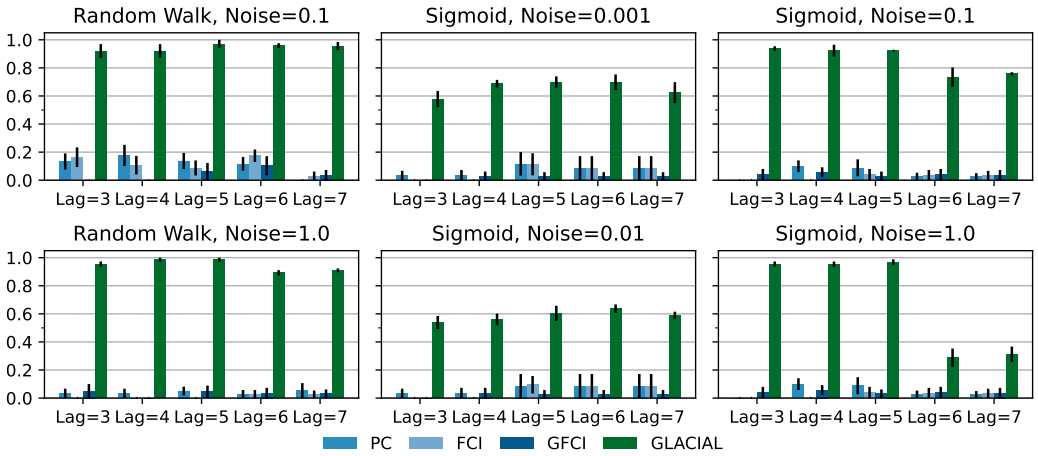


Figure 15: Performance of PC, FCI, and GFCI when thresholded at 0.8 (80%).

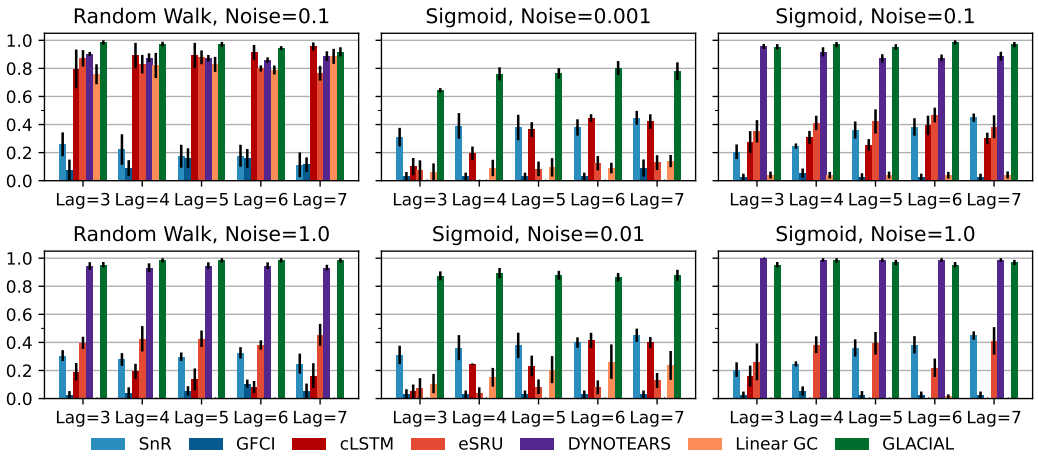


Figure 16: Average F1-scores at different settings of sample path, lag-time and measurement noise (7-node graph). Each subject has 24 timepoints. GLACIAL outperforms baselines in most settings.

Appendix D. Experiment on ADNI Dataset

D.1 Description of ADNI Variables

Table 2 shows the ADNI variables used and how they were measured (Data Modality). These variables are complementary in what they measure. "ABETA" and "PTAU" measure the level of two proteins in cerebral spinal fluid that are indicative of Alzheimer's disease. "FDG" measures brain cells' metabolism while cognitive tests measure performance in various areas such as general cognition, memory, language et cetera. The quantitative variables derived from structural MRI scans (e.g. Hippocampus Volume) is often considered

as a proxy of regional brain atrophy, or tissue loss linked to aging and/or neuro-degenerative processes.

Table 2: ADNI variables used for causal discovery

Variable	Description	Data Modality
ABETA	Amyloid beta	Cerebral spinal fluid
FDG	Fluorodeoxyglucose PET	PET imaging
PTAU	Phosphorylated tau	Cerebral spinal fluid
ADAS13	ADAS-Cog13	Cognitive test
MMSE	Mini-Mental State Examination	Cognitive test
MOCA	Montreal Cognitive Assessment	Cognitive test
Entorhinal	Entorhinal cortical volume	MRI imaging
Fusiform	Fusiform cortical volume	MRI imaging
Hippocampus	Hippocampus volume	MRI imaging
MidTemp	Middle temporal cortical volume	MRI imaging
Ventricles	Ventricles volume	MRI imaging
WholeBrain	Whole brain volume	MRI imaging

We normalized the volumetric variables of each subject by dividing the measurements by the subject's intracranial volume, or total head size, which is typically constant in adulthood. This is a standard normalization done to account for inter-subject variability in head sizes. FDG is a standardized uptake value ratio computed by dividing the average PET signal in an Alzheimer implicated region of interest to the signal in a control reference region. The Cerebral Spinal fluid markers correlate with the accumulation of the two Alzheimer's associated pathological proteins in the brain, namely tau tangles and amyloid plaque.

D.2 Interpretation of ADNI Causal Graphs

The order of the volumetric variables in Fig 7a, 7b, and 7c are mostly consistent with each other and prior literature on neuroimaging in aging and Alzheimer's disease, where the size of ventricles and whole brain are earliest MRI markers of aging, and Alzheimer's associated atrophy starts at the hippocampus, from where it spreads to cortical areas such as entorhinal and fusiform.The causal chains that appear in all three graphs are:

- "Ventricles" → "WholeBrain" → "Hippocampus" → "Entorhinal" → "Fusiform"
- "Ventricles" → "WholeBrain" → "MidTemp" → "Fusiform"

The ordering of cognitive tests are also consistent across all graphs. When we examine the ordering of variables from different data modalities, the causal chain of "Hippocampus" → "ADAS13" → "MMSE" → "MOCA" is quite interesting. This implies that the atrophy of the hippocampus, a brain region that plays a central role in memory and learning, leads to worse performance in tasks measured by cognitive tests. The relationship between MMSE and ADAS13 is surprising and deserves follow-up investigation, because classically MMSE is thought of as the earlier marker of cognitive impairment and ADAS13 is a measure of symptoms that appear later in Alzheimer's disease. That said, to our knowledge, we are

not aware of a study that examines the relationships between the temporal dynamics of these test scores. Our results indicate that changes in ADAS13 might foreshadow changes in MMSE.

Fig 17 shows the outputs of the Sort-N-Regress baseline and the remaining timeseries baselines on the ADNI data. The output of Sort-N-Regress seems implausible because the nodes FDG, PTAU, and ABETA are children of the MRI nodes. This contradicts the established literature about Alzheimer's disease in which indicates that the disease seems to originates first from changes in FDG, PTAU, and ABETA (hence these nodes should be roots instead of descendants). The outputs from PCMCI+ and TCDF are not very informative as they lack edges between the ROIs. Although somewhat similar GLACIAL's output, output from SVAR-GFCI has a lot of bidirectional edges. The outputs of SRU and cMLP also contain bidirectional edges. One of the possible reasons for the seemingly worse performance of the baselines is the high missing rate in the ADNI data.

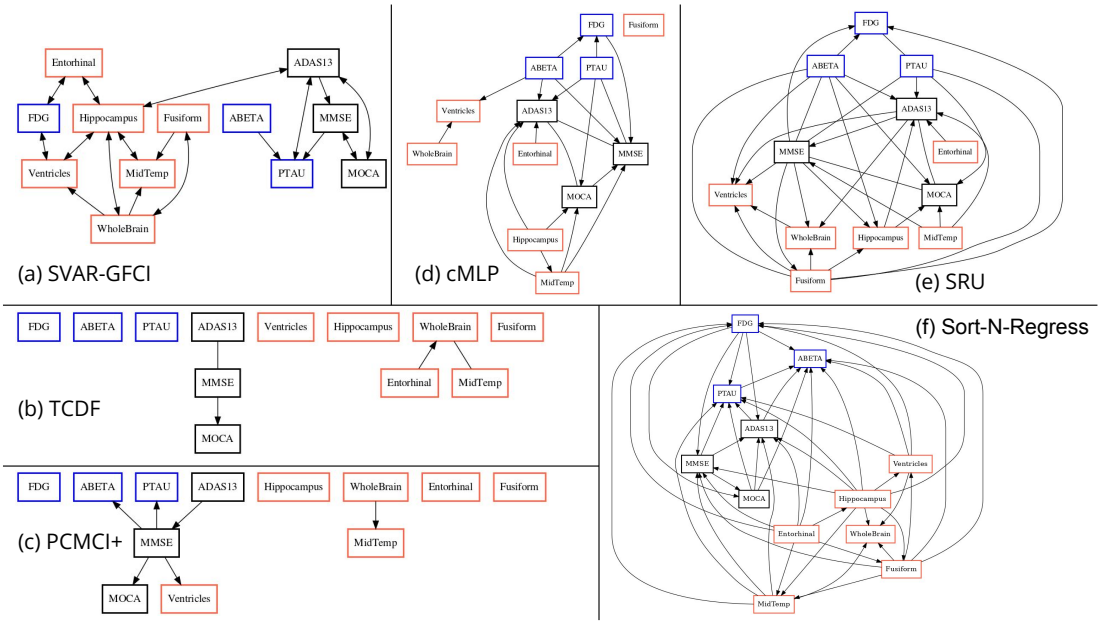


Figure 17: Baseline methods' predicted interaction of ADNI biomarkers. ROI volumes are in red, cognitive tests are in black, and the rest are in blue. ABETA: amyloid beta, PTAU: phosphorylated tau.