GRANGER CAUSAL INFERENCE ON DAGS IDENTIFIES GENOMIC LOCI REGULATING TRANSCRIPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental challenge in studying complex dynamical systems is determining causal relationships between a system's variables. When the system can be modeled as a time series of observations, Granger causality has proven to be a powerful approach for detecting interactions between its variables. Unfortunately, many domains do not lend themselves to a total ordering of observations, e.g., cell differentiation trajectories, citation networks, Twitter retweet graphs etc. Instead, the dynamics in such domains are more naturally represented by directed acyclic graphs (DAGs) in which nodes represent distinct dynamic states and information flows along directed edges between nodes. Here, we introduce the problem of detecting Granger causal interactions for such DAG-structured dynamical systems while also accounting for potential nonlinearity in the interactions. While our problem formulation is general and broadly applicable, our motivating application is the analysis of single-cell gene regulatory dynamics for identifying genomic loci that participate in the regulation of specific genes. We propose GrID-Net, a novel framework based on graph neural networks for Granger causal inference that captures long-range nonlinear predictive relationships between variables along a DAG. Applying it on multimodal single-cell assays that profile chromatin accessibility and gene expression in the same cell, we show that GrID-Net substantially outperforms existing methods for inferring regulatory region–target gene links. We also use GrID-Net to detect topic–journal relationships in a citation network of 15,566 publications. GrID-Net thus offers the potential to generalize the application of Granger causality to new domains and, in particular, reveal the regulatory interactions relevant to cellular differentiation and human diseases.

1 INTRODUCTION

Understanding the structure of a multivariate dynamical system often boils down to deciphering the causal relationships between its variables. Since inferring true causality is often infeasible, requiring additional experiments or independent mechanistic insight, statistical analysis of observational data that identifies predictive relationships between the system's variables can be very valuable. Specifically, the framework of Granger causality enables us to study a dynamic system by quantifying pairwise interactions between its stochastic processes (Granger, 1969; Shojaie & Fox, 2021). Intuitively, in a dataset where observations are temporally ordered, a time-dependent variable x (with value x_t at time t) is said to "Granger cause" a second variable y if the history of x (i.e., $x_1, \ldots, x_{t-1}$) is significantly predictive of y_t even after accounting for y's own history. Originating in econometrics, Granger causality has been a powerful analytic tool in many domains including finance, social sciences, and climatology (Benhmad, 2012; Rasheed & Tahir, 2012; Zhang et al., 2011). In a biological context, it has been used for gene regulatory network inference from time-resolved bulk transcriptomic assays and neural activity mapping (Fujita et al., 2007; Yao et al., 2015; Roebroeck et al., 2005). A prerequisite for applying Granger causality, however, is that the data must conform to a *total* ordering along time, i.e., there be a clear sequential ordering of observations.

Often, a *partial* ordering of the observations is more natural. In some social network settings, information flows along a tree (e.g., the retweets of a meme by the followers of an earlier retweeter). In text-mining, the citation graph of publications captures the flow of scientific ideas and knowledge. The set of bitcoin transactions in the blockchain ledger is also partially ordered. In all such cases, the dynamics of the system are more suitably represented as a directed acyclic graph (DAG) corre-

sponding to the partial ordering: nodes in the DAG represent distinct dynamical states and its edges indicate the flow of information between states. We limit our focus here to *strict* partial orderings, noting that each such ordering corresponds to a DAG (and vice-versa).

Our motivating example is from single-cell genomics where the partial ordering of observations and its associated DAG arise from the experimental setup. There, transcriptomic and/or epigenomic features are profiled per cell across a population of cells, in order to understand the dynamics of cellular processes (Gawad et al., 2016; Buenrostro et al., 2015). Since cells are destroyed during measurement, the observed per-cell data only represent snapshots of the underlying dynamics and are not time-resolved. While computational techniques can approximate a *pseudotime* ordering of the cells based on the similarity of their profiled features (Haghverdi et al., 2016), forcing this richly heterogeneous dataset into a sequential ordering of cells along pseudotime can be suboptimal due to the numerous sources of biological and technical variation in single-cell data (Zhang et al., 2019) as well as multiple cells representing the same cellular state (Baran et al., 2019). Instead, a more appropriate representation of the data involves considering the entire manifold of cellular states, typically represented by a k-nearest neighbor (kNN) graph of cells. Orienting the edges of the kNN graph to agree with the estimated pseudotime ordering of cells then yields a DAG that captures the diversity of cell states as well as their dynamics.

The key conceptual advance of this work is extending the applicability of Granger causality to systems with partially ordered observations. On a DAG where each node takes on multivariate values, we define a variable x as Granger causing another variable y if the latter's value y_v at the DAG node v can be significantly predicted by the values of x at the ancestor nodes of v, even after accounting for y's own ancestral history. We introduce GrID-Net (**Gr**anger **I**nference on **D**AGs), a graph neural network (GNN) framework for Granger causal inference in dynamical systems that can be represented as DAGs (Figure 1a). We modify the standard GNN architecture to enable lagged message-passing on the DAG, enabling us to accumulate past information for a variable according to the DAG's structure. This, combined with a mathematical formulation of Granger causality more amenable to neural networks, allows GrID-Net to recover nonlinear long-range Granger causal dependencies between the variables.

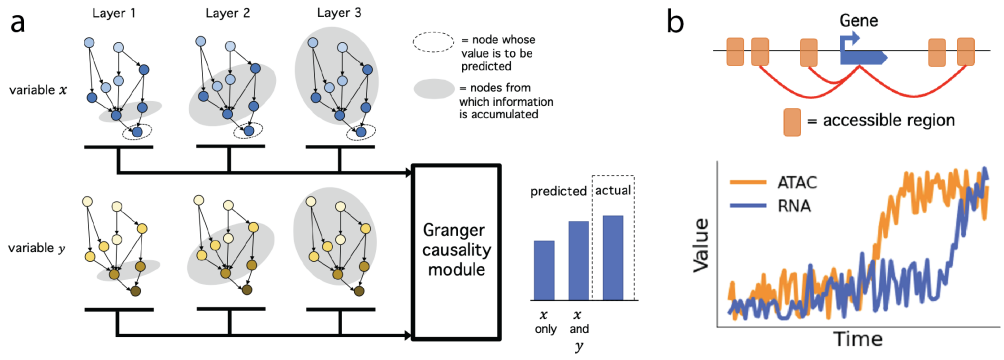

Figure 1: (a) Schematic of the GrID-Net framework. The

We apply GrID-Net on data from single-cell multimodal assays that profile chromatin accessibility and gene expression in the same cell. By simultaneously identifying genomic regions that are accessible for gene regulators to bind as well as measuring each gene's expression, these datasets promise the ability to detect cell type-specific links between noncoding genomic sites and the genes they regulate (Figure 1b). Current methods for analyzing this data seek to correlate the contemporaneous measurements of the two modalities. However, the two modalities are not temporally synchronized: chromatin accessibility precedes and has a causal influence on gene expression (Ma et al., 2020; Chen et al., 2019). In dynamic processes such as differentiation and immune cell activation, the priming of chromatin state has been observed (Lara-Astiaso et al., 2014; Ostuni et al., 2013), with the activation of a specific chromatin state preceding the downstream gene expression changes it triggers. Hence, an approach that can recover these predictive relationships is needed.

We show that GrID-Net substantially outperforms current methods in identifying known regulatory region-target gene links in three single-cell multimodal datasets characterizing a range of dynamic

processes, including cancer drug responses and cellular differentiation. We also demonstrate that GrID-Net yields strong empirical results in other domains, including text-mining of literature, suggesting the broad utility of this framework. To our knowledge, this work presents the first framework for Granger causal inference on dynamical systems represented as DAGs, the first to explicitly account for temporal asynchrony between chromatin accessibility and gene expression, and opens up opportunities to uncover new insights from the abundant and growing collection of graph-based datasets.

2 METHODS

In this section, we provide an overview of Granger causality and then describe the components of our framework in detail.

2.1 GRANGER CAUSALITY: STANDARD FORMULATIONS

In time series analysis, Granger causality is typically evaluated using the vector autoregressive (VAR) model, which can be expressed as follows (Lütkepohl, 2005).

$$y_t = \sum_{\ell=1}^{L} a_1^{(\ell)} y_{t-\ell} + \sum_{\ell=1}^{L} a_2^{(\ell)} x_{t-\ell} + \epsilon_t \tag{1}$$

Here, $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}$ are values at time t for two stationary time series y and x, respectively; $a_1^{(\ell)}$ and $a_2^{(\ell)}$ are coefficients that specify how lag ℓ affects the future of time series y; and ϵ_t is a zero-mean noise term. In this model, time series y is assumed to be a linear combination of the past L lags of y and x each. Time series x is said to *not* "Granger cause" time series y if and only if $a_2^{(\ell)} = 0$ for all ℓ.

Another formulation— the one we follow— for performing Granger causal inference for generalized models is to consider two related models for forecasting time series y. A *full* model considers the past values of both y and x to forecast y, while a *reduced* model excludes the effect of x, only containing terms related to the past values of y.

$$y_t = f^{(full)}(y_{t-1}, ..., y_L; x_{t-1}, ..., x_{t-L}) + \epsilon_t \tag{2}$$

$$y_t = f^{(reduced)}(y_{t-1}, ..., y_{t-L}) + \epsilon_t \tag{3}$$

Here, $f^{(full)}(\cdot)$ and $f^{(reduced)}(\cdot)$ are generalized functions that specify how the value of y_t depends on past values of y and x. The predictions of each of these models are then compared, upon which a Granger causal relationship is declared if the full model's predictions are significantly more accurate than those of the reduced model.

2.2 GRANGER CAUSALITY ON A DAG: GRAPH NEURAL NETWORK FORMULATION

Let the data be represented by a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ observations and the directed edges $\mathcal{E}$ indicating the information flow or partial order between these observations. Let $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{x} \in \mathbb{R}^n$ correspond to the values of variables y and x on the nodes in $\mathcal{V}$, with x putatively causing y. To infer Granger causality, we compare full and reduced models as in Eqns. 2,3:

$$\boldsymbol{y} = g(\tilde{\boldsymbol{h}}_y^{(full)} + c\tilde{\boldsymbol{h}}_x^{(full)}) + \boldsymbol{\epsilon} \tag{4}$$

$$\boldsymbol{y} = g(\tilde{\boldsymbol{h}}_y^{(reduced)}) + \boldsymbol{\epsilon} \tag{5}$$

Here, $\tilde{\boldsymbol{h}}_y^{(full)}, \tilde{\boldsymbol{h}}_y^{(reduced)}, \tilde{\boldsymbol{h}}_x^{(full)} \in \mathbb{R}^n$ represent the history of y or x (as denoted by the subscript) for each of the n observations, i.e., the information accumulated from their respective ancestors in $\mathcal{G}$. c is a coefficient that describes the effect of the history of variable x on the value of variable y. In addition, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a zero-mean noise term. We also include $g(\cdot)$ as an optional link function that maps its input to the support of the distribution of variable y.

For brevity, we only describe how $\tilde{\boldsymbol{h}}_x^{(full)}$ is computed, with $\tilde{\boldsymbol{h}}_y^{(full)}$ and $\tilde{\boldsymbol{h}}_y^{(reduced)}$ being computed analogously. Also, since $\boldsymbol{x}$ occurs only in the full model, for notational convenience we drop its superscript *(full)* below, writing just $\tilde{\boldsymbol{h}}_x$. We express $\tilde{\boldsymbol{h}}_x$ as the mean of the outputs of L consecutive graph neural network (GNN) layers, denoting the layerwise outputs as $\boldsymbol{h}_x^{(\ell)}$:

$$\tilde{\boldsymbol{h}}_x = \frac{1}{L}\sum_{\ell=1}^{L} \boldsymbol{h}_x^{(\ell)} \tag{6}$$

$$\boldsymbol{h}_x^{(\ell)} = \begin{cases} f_x^{(\ell)}(\boldsymbol{h}_x^{(\ell-1)}, \boldsymbol{A}_+) & \text{if } \ell > 1 \\ f_x^{(\ell)}(\boldsymbol{x}, \boldsymbol{A}) & \text{if } \ell = 1 \end{cases} \tag{7}$$

Here, $\boldsymbol{A}$ and $\boldsymbol{A}_+ \in \mathbb{R}^{n\times n}$ are matrices defined by the DAG structure, designed to introduce a one-step lag in the first GNN layer that is propagated by subsequent layers. $\boldsymbol{A}$ is the column-normalized adjacency matrix of $\mathcal{G}$, with $A_{ij} = \frac{1}{d_j}$ if edge $(i,j) \in \mathcal{E}$ and 0 otherwise. d_j is the in-degree of node j, ensuring the column normalization. We note that a DAG does not have self-loops so that the diagonal terms of $\boldsymbol{A}$ are zero. In $\boldsymbol{A}_+$ we introduce these diagonal terms, setting them to the same value as others in the column (d_j is incremented by 1 to preserve normalization).

Together, $\boldsymbol{A}$ and $\boldsymbol{A}_+$ allow us to capture the key intuition of Granger causality on a DAG: predicting a value at node v from the values of its ancestors that are within L steps. The first layer introduces the lag, using information at the parents of v but not the information at v itself. Each subsequent layer introduces the preceding set of parents, with the diagonal term in $\boldsymbol{A}_+$ ensuring that information stored at v from previous layers is integrated with the new information propagated via v's parents. The sequence of GNN layers therefore reflects the successive aggregation of past information by traversing L steps backward along the DAG for each node in the graph. Thus, we are able to match the concept of lags in time series-based formulations (e.g., Eqn. 1) but with the crucial advantage of leveraging the richness of a DAG's structure in directing the information flow.

We compute the GNN layers as follows (again, we only describe f_x for brevity):

$$f_x^{(\ell)}(\boldsymbol{h}_x^{(\ell-1)}, \boldsymbol{A}_+) = \sigma(w_x^{(\ell)} \boldsymbol{A}_+^T \boldsymbol{h}_x^{(\ell-1)} + b_x^{(\ell)}) \quad \text{if } \ell > 1 \tag{8}$$

$$f_x^{(\ell)}(\boldsymbol{x}, \boldsymbol{A}) = \sigma(w_x^{(\ell)} \boldsymbol{A}^T \boldsymbol{x} + b_x^{(\ell)}) \quad \text{if } \ell = 1 \tag{9}$$

where $w_x^{(\ell)}, b_x^{(\ell)} \in \mathbb{R}$ are the per-layer weight and bias terms, respectively. $\sigma(\cdot)$ represents the nonlinear activation in each of the graph neural network layers, which we have chosen to be the hyperbolic tangent function *tanh*$(\cdot)$.

2.3 Comparing Full and Reduced Models to Infer Granger Causality

Let $\Theta_y^{(full)}, \Theta_y^{(reduced)}, \Theta_x^{(full)} \in \mathbb{R}^{2L}$ denote the set of parameters for the full and reduced models, with $\Theta_x^{(full)} = \{w_x^{(1)}, ..., w_x^{(L)}; b_x^{(1)}, ..., b_x^{(L)}\}$ and $\Theta_y^{(full)}, \Theta_y^{(reduced)}$ defined analogously. We propose to jointly learn all the Θ parameters by minimizing the following loss function where we evaluate the full and reduced models simultaneously:

$$\mathcal{L}^{(total)} = \mathcal{L}^{(full)} + \mathcal{L}^{(reduced)} = \sum_{x,y\in\mathcal{P}} \left(\sum_{n=1}^{N} \mathcal{L}(\hat{y}_n^{(full)}, y_n) + \sum_{n=1}^{N} \mathcal{L}(\hat{y}_n^{(reduced)}, y_n) \right) \tag{10}$$

In some multivariate systems (e.g., single-cell genomics) there may be with hundreds of thousands of choices for x or y, not all pairwise combinations of these variables might be of interest. Above, $\mathcal{P}$ represents the set of x, y variable pairs whose putative causal interactions is of interest. $\hat{y}_n^{(full)}$ and $\hat{y}_n^{(reduced)}$ correspond to the predictions for observation n by the full and reduced models, respectively.

To infer Granger causal interactions between a pair of variables y and x, we compare the set of loss terms associated with the full model $\mathcal{L}(\hat{y}_n^{(full)}, y_n)$ to the those of the reduced model

$\mathcal{L}(\hat{y}_n^{(reduced)}, y_n)$. We use a one-tailed Welch's t-test, where the alternative hypothesis is that the mean of the loss terms for the full model is less than the mean of the loss terms for the reduced model. We then score the interaction between each pair of variables as the $-\log_{10}(p)$-value associated with the Welch's t-test and rank the set of interactions with respect to these scores, with higher scores indicating a more significant Granger causal interaction.

2.4 Training Details

The graph neural network models are trained using the Adam optimizer with a learning of 0.001 for 20 epochs or until convergence, which we define to be the point at which the relative change in the loss function is less than $0.1/|\mathcal{P}|$ across consecutive epochs. A minibatch size of 1024 was used during training, and trainable parameters in the model were initialized using Glorot initialization (Bengio & Glorot, 2010). In our applications of GrID-Net to single-cell multimodal data, we chose the loss function to be the negative log likelihood under the Poisson likelihood model, while for the citation network application, we used the binary cross entropy loss. All models were implemented in PyTorch and trained on compute details here.

3 Related Work

lorem ipsum

4 Results

lorem ipsum

4.1 Chromatin Accessibility

lorem ipsum

4.2 Citation Network

lorem ipsum

5 Discussion

lorem ipsum

Author Contributions

lorem ipsum

Acknowledgments

lorem ipsum

References

Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. *Genome Biology*, 20(206), 2019.

Yoshua Bengio and Xavier Glorot. Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of AISTATS 2010*, 9:249–256, 2010.

François Benhmad. Modeling nonlinear Granger causality between the oil price and US dollar: A wavelet based approach. *Economic modelling*, 29(4):1505–1514, 2012.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 18(7561):486–490, 2015.

Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37:1452–1457, 2019.

Andr Fujita, Joo R. Sato, Miguel Garay, Rui Yamaguchi, Satoru Miyano, Mari Sogayar, and Carlos E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(39), 2007.

Charles Gawad, Winston Koh, and Stephen R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17:175–188, 2016.

Clive W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016.

David Lara-Astiaso, Assaf Weiner, Erika Lorenzo-Vivas, Irina Zaretsky, Diego Adhemar Jaitin, Eyal David, Hadas Keren-Shaul, Alexander Mildner, Deborah Winter, Steffen Jung, Nir Friedman, and Ido Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949, 2014. doi: 10.1126/science.1256271.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. MIT Press, 2005.

Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Law Travis Tay, Tristan, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116.E20, 2020.

Renato Ostuni, Viviana Piccolo, Iros Barozzi, Sara Polletti, Alberto Termanini, Silvia Bonifacio, Alessia Curina, Elena Prosperini, Serena Ghisletti, and Gioacchino Natoli. Latent enhancers activated by stimulation in differentiated cells. *Cell*, 1-2:157–171, 2013.

Hafsa Rasheed and Muhammad Tahir. Fdi and terrorism: Co-integration & granger causality. *International Affairs and Global Strategy*, 4(0):1–5, 2012.

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

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *arXiv preprint arXiv:2105.02675*, 2021.

Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC Bioinformatics*, 16(273), 2015.

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. *Proceedings of the National Academy of Sciences*, 108(42):17296–17301, 2011.

Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature Communications*, 10(2611), 2019.

A Appendix

You may include other additional sections here.