EPiC Series in Computing

Volume 83, 2022, Pages 10–19

Proceedings of 14th International Conference on Bioinformatics and Computational Biology

Learning Gene Regulatory Networks using Graph Granger Causality

Pranita Patil and Maria Vaida

Harrisburg University of Science and Technology, Harrisburg, US
pppatil@my.harrisburgu.edu, mlvaida@my.harrisburgu.edu

Abstract

Interacting systems such as gene regulatory networks have the ability to respond to individual component changes, propagate these changes throughout the network, and affect the temporal trajectories of other network elements. Causality techniques are frequently employed to investigate the interconnection between variables in complex dynamical systems. However, the vast majority of causality models are rooted in regression techniques such as Vector Autoregression Models and Bootstrap Elastic net regression from Time Series framework, and there is very limited research in the space of deep learning, particularly graph neural networks. In this paper, we explore in more depth the concept of Granger causality in deep learning and propose Granger causality deep learning framework using graphs convolutions, LSTM, and nonlinear penalties for the objective of learning causal relationships between temporal elements in gene regulatory networks. The deep learning architecture proposed here for studying causality in dynamic networks has achieved high results on simulated networks as well as on more challenging Dream3 gene regulatory networks time-series datasets.

1 Introduction

The advances in the field of mRNA sequencing have been opening the doors for a wide range of applications in the research and medical communities. mRNA sequencing and analysis hold the potential to unlock great insights into disease causes and progression, carrier status, cancer, and infectious diseases research, as well as other fields such as agrigenomics. Understanding the causes that lead to changes in gene regulatory expression levels over time and across conditions is vital in studying the genesis, progression, and ultimately treatments of diseases.

Gene regulatory network inference methods seek to uncover these complex relationships among gene pathways and predict the consequences of perturbations. Existing causality frameworks focus on studying the associations between elements in an interacting system using statistical quantitative techniques that require apriori assumptions on the data. Moreover, these models are not able to fully capture the nonlinearities hidden in the latent variables, and the vast amounts of available data that are idiosyncratic to the mRNA analysis can generate high levels of false positives. Data science on the other hand can overcome these limitations while learning latent structural properties in connected networks. More specifically, graph neural

networks have been proven an essential tool in studying networks, encoding and reconstructing structural characteristics of nodes, and predicting network linkage.

This work extends causality inference to graph neural networks and combines them with existing LSTM causality frameworks introduced by [20], with the aim of learning causal relation—ships from transcriptomal time series data based on Granger causality. Graph Granger Causality (GGC) framework proposed here also incorporates nonlinearities hidden in the data as weight initializers and model penalties in the unsupervised task of learning latent causal structure in temporal gene regulatory networks. The concept of node-node similarities in graph networks have been extensively used in graph neural networks by [7, 3, 12, 17]. In the field of causality, correlation-based similarity measures have been used by [9] and [4].

2 Related Work

Interacting biological systems such as neurons, proteins, and genes have a strong interdependence expressed through various mechanisms that propagate individual component perturbations across the entire network. However, the latent causal relationships between components of these systems are hidden. To understand network dynamics one must infer causal relations from the available time-based observational data [24]. Time series causality inference quantifies the degree to which one variable evolution in time impacts another variable trajectory. Such a causal relationship can be translated into the ability of the first variable to explain or predict the the second variable[23], [6]. A variable Z is considered Granger causal of variable X if the past values of $Z_{t,\ldots,t-1}$ improves the prediction accuracy of future values of X_{t+1} when compared with a model in which $Z_{t,\ldots,t-1}$ is not included.

In the gene regulatory networks (GRN) research, the Granger causality space has been mostly focused around regression techniques. For instance, Bootstrap Elastic net regression from Time Series (BETS) infers causal relationships in gene regulatory networks using elastic net regression and stability selection from bootstrapped samples. In addition to causal relationships prediction, BETS also infers the directionality of effect, namely whether causal effects are activating or inhibitory [11]. Sliding Window Inference for Network Generation (SWING) is another time series windowed network inference approach based on Granger causality. SWING identifies associations between genes by applying an ensemble of regression-based models over time-series gene expression data using flexible time-series windows [5]. For each time window, SWING infers a ranked list of time-delayed causal edges between variables.

The alternative techniques used for Granger causality are based on prior information and conditional probabilities such as transfer entropy [22] and models based on maximum likelihood estimation [13]. A more advanced method such as Gene regulatory networks on transfer entropy (GRNTE) is implemented to infer gene regulatory interactions by [2]. This method uses partial mutual information between pairs of genes and higher values of transfer entropy correspond to stronger interaction. Prior knowledge-driven Granger causality model [25] uses the prior knowledge as the directed or undirected weighted graph between the gene. This conditional Granger causality model also incorporates ridge regularization to resolve the problem of data size limitation.

More recently [24],[15] and [18] proposed causality frameworks that account for network nonlinearities. Large-scale nonlinear Granger causality (lsNGC) [24] models high dimensional limited-time series interval data with nonlinear dimensionality reduction techniques using radial basis functions to identify statistically significant casual relations. lsNGC is applied to functional Magnetic Resonance Imaging (fMRI) data to detect causality in brain tissues. Hilbert–Schmidt independence criterion Lasso Granger causality (HSIC-Lasso-GC) [15] extracts

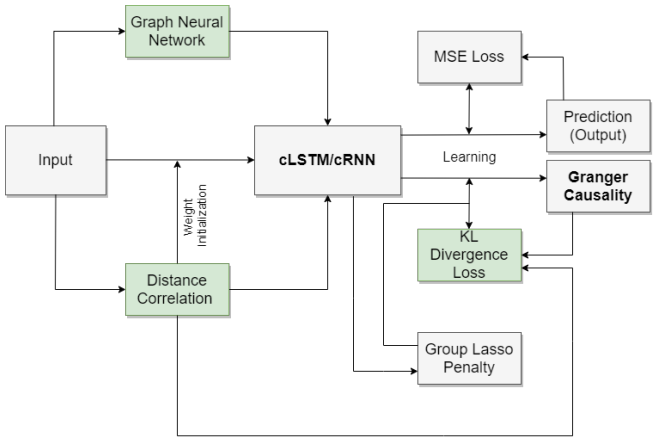

Figure 1: GGC Model per one predictor variable.

nonlinear time series intimation using stationarity test and state-space reconstruction functions and applies a HSIC-Lasso model to learn the causality structure. Backward-in-Time Selection Conditional Granger Causality index (BTS-CGCI), defines Granger causality as the logarithm of the ratio of the error variances of leave-one-variable-out and the unrestricted ordinary least squares (OLS) models. BTS-CGCI incorporates the assumption that the variables at smaller lags are more explanatory to the response variable than variables at larger lags. Thus conditioning on the variables at smaller lags the variables at larger lags may only enter the model if they have genuine contribution not already contained in the selected variables of smaller lags [18].

All previous models mentioned above either face problem of a data size limitation or are not able to capture linear dependencies or curse of dimensionality. Tank, A. et all. were the first to introduce a time series causality inference deep learning architecture. Their model titled Neural Granger Causality (NGC) combines fully connected layers and long-short term memory convolutions (LSTM) with group-lasso penalties to extract the Granger causal structure from network parameters without any supervised causality loss measures [20]. Our models build on this deep learning framework by integrating graph neural networks with LSTM convolutions to better capture network structure, connectivity, and time series interaction patterns in gene regulatory networks. We also extend the notion of penalties to include nonlinear similarities between variables, and to offer the model data-driven weights initialization. The framework of our model is shown in Figure 1.

3 Methodology

Graph Granger Causality (GGC) model combines graph neural networks over temporal data with Granger causality principle to measure the causal effects in a series of data points. Given a model with N variable of size p and time points T, Granger causality defines a causal relation between a pair of variables $N_{1i},...,_{Ti}$ and $N_{1j},...,_{Tj}$ by minimizing the reconstruction error of $M_{N_1,..,N_i,..,N_p}$ models. The model applies LSTM and graph convolutions over a set of variables expressed as a sequence of time points, by learning to reconstruct the sliding window in the time series.

Each model corresponds to one output variable, thus the number of variables $N_{1,...,p}$ dictates the total number of models $M_{N_{1,...,p}}$ the network will learn over as in [20]. Since recurrent models perform impressively for modeling time series data even with longer time series dependencies, we applied component-wise models with graph neural network to the recurrent neural network (RNN) and long-short term memory network (LSTM) and term those as Graph Granger Causality-component wise models such as GGC-cRNN and GGC-CLSTM models respectively. For each variable, the network convolves over the entire time-series and graph datasets and outputs the reconstruction values for one variable.

We also implemented a leave-one-out variation of the model (GGC-LOO), which includes an additional global model, thus in total the network having $M_{N_{1,...,p}} + 1$ models. The global model computes the loss over all variables and predicts each variable, while the remaining M models masks each of the $N_1, ..., p$ variable at a time to predict each variable. If we train separate models for all variables except jth variable to predict each variable i and the loss of the global model is less than the loss for jth variable N_j of the masked models, then we can conclude that the masked variable N_j is causal of N_i since its presence in the network reduces the reconstruction error. The input to this model is two streams of data: a time series dataset and a graph adjacency matrix of the variables.

The graph dataset is defined as a weighted undirected graph $\mathrm{G} = (\mathcal{V},\mathrm{E_w})$, where $\mathcal{V} = \{\upsilon_1, ..., \upsilon_p\}$ is a set of vertices, and $\mathrm{E}_w = e_{\imath j}$ represents the set of edges. Each node in the graph $\upsilon_\imath \in \mathcal{V}$ corresponds to a gene and each weighted edge encodes the normalized gene expression values. The structure of the graph is represented by its symmetric adjacency matrix $\mathrm{A} \in \mathrm{R}^{p\times p}$, where $\mathrm{A}_{\imath j} = W$, ($W\imath j$ is the edge weight between vertices), if the two genes, $\imath j$, are expressed together otherwise $\mathrm{A}_{\imath j} = 0$.

The normalized adjacency matrix is defined as a weighted Laplacian adjacency matrix derived from the time-series incidence matrix as follows:

$$\begin{aligned} \mathrm{A} &= N^T N \\ \mathrm{A}_{\imath j} &= \sum_{v=1}^{T} N_{iv} N_{vj} \text{ where } i,\, j = 1, .., p \end{aligned} \tag{1}$$

where $\sum_{\mathrm{v=1}}^{\mathrm{T}}$ represents the weights of the edges i and j if there is a common vertex v incident these edges. The edges are aggregated over all time points T . The resulting adjacency matrix values are then normalized.

The models consists of 3 layers: LSTM layer, GNN layer, and a Conv1D unit. The LSTM layer initiates its initial hidden states to the distance correlation of the data, which measures the nonlinear dependencies between variables and is computed as:

$$\mathcal{C}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}} & \text{if } \mathcal{V}^2(x,x)\mathcal{V}^2(y,y) > 0 \\ 0 & \text{else } 0 \end{cases} \tag{2}$$

By taking the square root of $C^2(X,Y)$, distance correlation $C(X,Y)$ is calculated. $C(X,Y)$ is between 0 and 1 when the variables have a level of interdependence, and $C(X,Y) = 0$ when the variables are independent. $v^2(x,y)$ is the distance covariance between a pair of variables as defined in [19]. All other LSTM components are as per [16].

Graph convolution layer integrates the structural properties of the network with the LSTM embeddings, which are then passed to the Conv1D layer. This last layer concatenates the

embeddings of all variables into one vector (which is representative of the gene time-series) the respective model is being optimized over. The complete model takes the following form:

$$h_i^{t(k)} = \sum_{j=1}^{p} \mathrm{A}_{ij}\ \mathrm{h}_j^{t(k-1)} \text{ where } i = 1, .., p \tag{3}$$

where $h_i^{t(k)}$ is the k window of time t output for i gene in the network, $\mathrm{h}_j^{t(k-1)}$ is the LSTM hidden state for time t window of $k-1$, A_{ij} is the graph adjacency matrix described in equation 1, and $\sum_{j=1}^{p}$ is summation over all variables which is nothing but genes. This last layer is convolution layer which condenses the dimensionality of the data from N variables to 1 variable window series that is being reconstructed for each k window.

In the optimization task, the loss function applies to mean squared error(MSE) between network predictions and output values for each time-series window and for each variable. Similar to [20], we optimize the models simultaneously by adding group lasso penalty. To further optimize the networks to learn the latent Granger structure, we add two additional loss constraints in the form of the distance correlation matrix described in equation 2, and Triplet loss [21]. Distance correlation also serves as the LSTM hidden input weights initialization as discussed earlier. Triplet loss minimizes the distance between similar samples and increases the distance between negative and positive samples. The complete loss function is computed as:

$$L = \min_{w} \sum_{t=2}^{T} \left(y_t^k - \hat{y}_t^k\right)^2 + \lambda_1 \sum_{j=1}^{p} ||W_1^j||_2^2 + \lambda_2 \sum_{n=1}^{p} y_n \cdot (\log y_n - x_n) + max(d(a, po) - d(a, ne) + R) \tag{4}$$

where $\sum \left(y_t^k - \hat{y}_t^k\right)^2$ is the MSE loss, $\lambda \sum ||W_j^i||_2^2$ is the Lasso penalty, $(\log_n - x_n)$ is the Kullback-Leibler divergence between the distance correlation and the reconstructed latent Granger causality matrix. $Max(d(a, po) - d(a, ne) + R)$ defines the Triplet loss, with $d(x_i, y_i = ||x_i - y_i||_p$ measuring the relative similarity between the anchor a, positive examples po and negative examples ne, and $R = 1$ is the soft margin distance between positive and negative examples. λ_1 is regularization parameters and controls the sparsity of Granger causal connections whereas λ_2 controls tradoff between observed and predicted non-linear dependence.

The time-series dataset for these proposed models is segmented into p X p sliding windows with a lag of 1 timepoint for each window. For each variable in the dataset, a model convolves over all variables and all windows and outputs the reconstruction time-series values for the next window in the series.

4 Experimental Results

We conducted experiments on 1 simulated Lorenz-96 dataset [10] and 5 gene regulatory network datasets that were published as part of the Dream3 challenges [1]. The Lorenz-96 model is used to simulate non-linear time series data with p=10 and data time series length, T=1000. The forcing constant used is 10. The Dream3 simulated data mimics gene expression and regulation dynamics, encoding latent nonlinear causal relationships, as described in [20]. The Dream challenge datasets include three sets of data, each with two E. Coli data sets and three Yeast networks. This dataset includes 10 and 100 genes spanning over 966 timepoints each and it is specifically designed to be a more challenging non-linear dataset [14].

The weighted adjacency matrices of the 5 gene networks and 1 simulated dataset are shown in Figures 3, 4, 15, 9, 10, and 18. By looking at these adjacency matrices A_{10x10}, we can interpret that if there is an edge or not from node N_i to N_j. If value is zero that means there is no edge connecting N_i and N_j. Convolution operators compute time series embeddings by infusing the global structural properties of each graph encoded in the adjacency matrices into the calculations. The graphical representations 5, 7, 16, 11,13, and 19 are force-directed drawing which applies the ForceAtlas2 algorithm to learn the connections between nodes in order to create a structural map of the network. The generated visual graphs can be interpreted as structural similarity densities. The closer the nodes are, the weaker the attraction force is, subject to the edge weights factor. The repulsion force decreases when distances increase [8]. Figures 5, 7, 16, 11,13, and 19 display network weight initial values modeled as the distance correlation plots derived from the time series data. These figures are distance correlation matrix C_{10x10} where each row and column corresponds to a gene. Higher values closer to 1 represents dependence between different genes whereas lower values closer to 0 denote genes are independent. As with the graphical data, distance correlation values capture similar dynamics between variables across time points.

We implemented GGC-cLSTM, GGC-cRNN, and GGC-LOO with 10 hidden units to detect complex non-linear dependencies in Dream3 10-gene datasets and Lorenz-96 datasets and compared our GGC approach with previously published results of models GC-cLSTM (Granger Causal-cLSTM) and GC-cRNN (Granger Causal-cRNN) [20]. accuracies for all models for all datasets are shown in Table 1. In terms of accuracy, our models GGC-cLSTM outperforms GC-cLSTM and GC-cRNN across all six datasets by a wide margin. GGC-cRNN follows a similar architecture with GGC-cLSTM, replacing the LSTM layer with a Recurrent Neural network (RNN) layer. This model outperformed all other models on the Yeast2 data (Table 1). These results indicate that including graphical interactions and non-linear dependencies as input to GGC-cLSTM networks boost the performance in recovering the interpretable non-linear interactions. The reason behind GGC-LOO not performing well might be the presence of a masked variable in inter-dependencies between remaining variables which are not masked.

In Figures 20 and 21, we also compute the AUROC and AUPR for all models across all five Dream 10-gene datasets since these performance metrics are commonly used for Dream datasets. As expected, the AUROC plot and the AUPR plot indicate that the GGC-cLSTM and GGC-cRNN perform better than the other models. Except for the yeast-2 dataset where GGC-cRNN slightly underperforms GC-cRNN.

To check the performance of GGC-cLSTM model on a higher dimensional time series dataset with a larger number of hidden units, we implemented 100 hidden units for the GGC-cLSTM for five 100-gene Dream Dataset. The results in Table 2 reveals that a higher number of hidden units improves the performance of GCC-cLSTM across all five 100-gene datasets as compared to 10-gene datasets. We also observed the computational cost of models increases with the number of variables with the 10-variable models being the most efficient.

5 Conclusions

Graph Granger Causality framework discussed here combines LSTM and graph convolutions with the objective of learning latent causal relationships in gene regulatory networks. Causal dynamics in networks are oftentimes rooted in interaction patterns that can be captured through various nonlinear functions such as distance correlation metrics. GGC captures these intra-gene nonlinearities through initialization and penalties strategies. First, the input weights of the LSTM layers are given more informative initialization values by sharing them with the

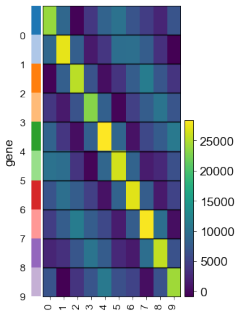

Figure 2: Weighted Lorenz adjacency matrix.

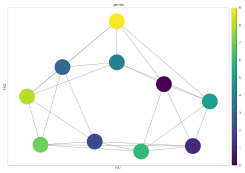

Figure 4: Lorenz graph.

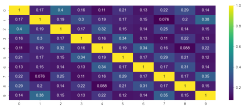

Figure 5: Lorenz Distance Correlation.

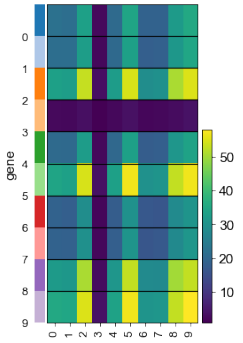

Figure 3: Weighted Yeast1 adjacency matrix.

Figure 6: Yeast1 graph.

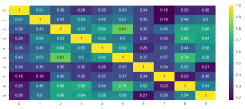

Figure 7: Yeast1 Distance Correlation.

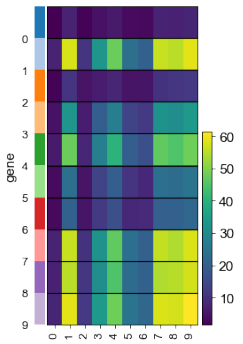

Figure 14: Weighted Ecoli1 adjacency matrix.

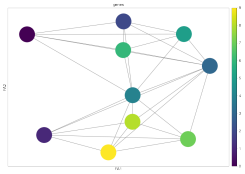

Figure 15: Ecoli1 graph.

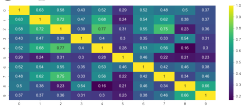

Figure 16: Ecoli1 Distance Correlation.

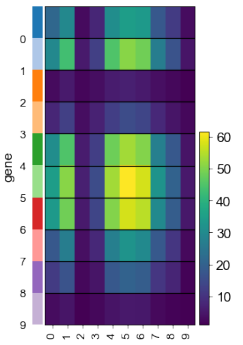

Figure 8: Weighted Yeast2 adjacency matrix.

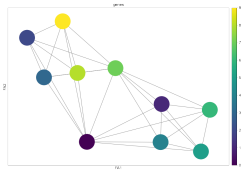

Figure 10: Yeast2 graph.

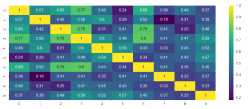

Figure 11: Yeast2 Distance Correlation.

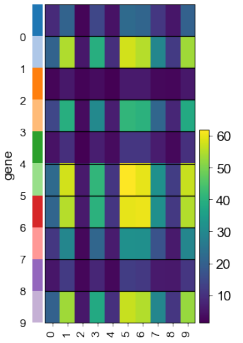

Figure 9: Weighted Yeast3 adjacency matrix.

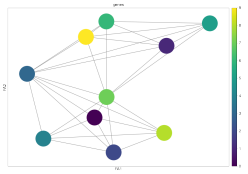

Figure 12: Yeast3 graph.

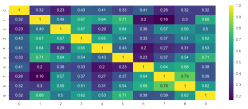

Figure 13: Yeast3 Distance Correlation.

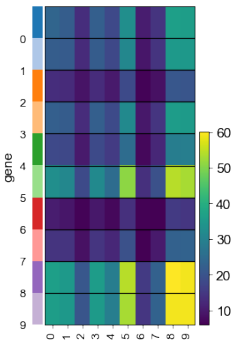

Figure 17: Weighted Ecoli2 adjacency matrix.

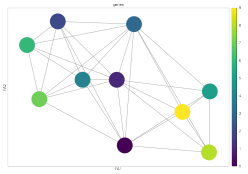

Figure 18: Ecoli2 graph.

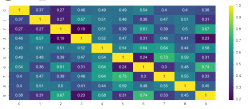

Figure 19: Ecoli2 Distance Correlation.

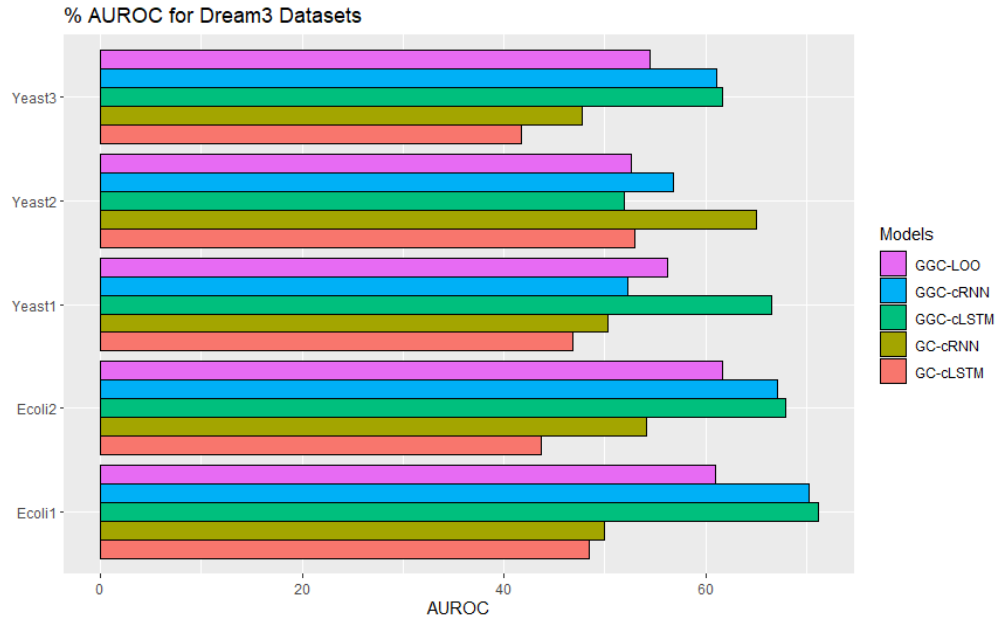

Figure 20: AUROC in percentage for Dream3 10-gene Datasets.

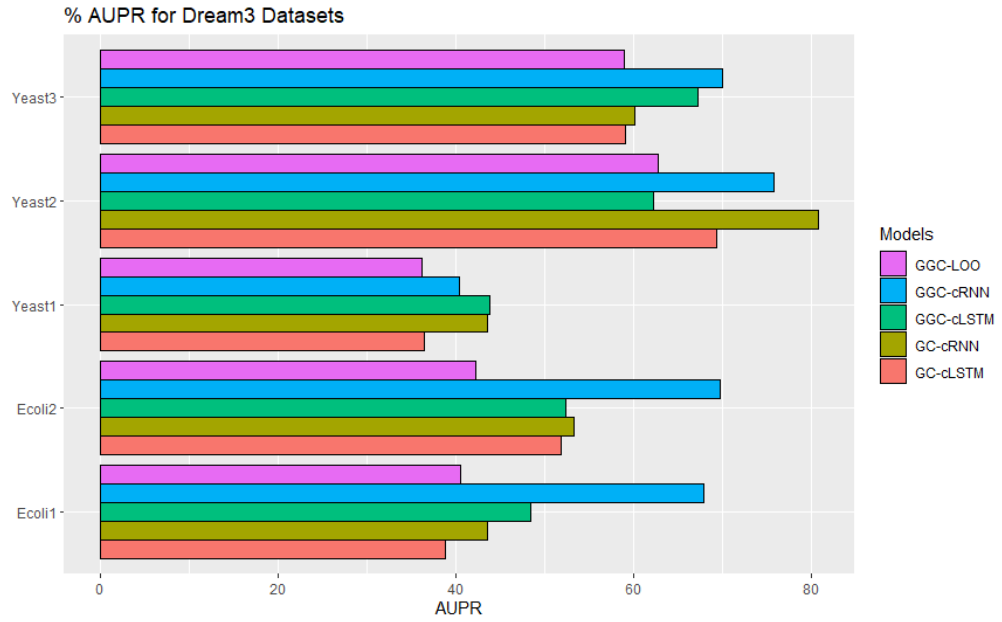

Figure 21: AUPR in percentage for Dream3 10-gene Datasets.

Model	Lorenz	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
GGC-cLSTM	**.99**	**.70**	**.64**	**.71**	.56	**.60**
GC-cLSTM	.97	.56	.42	.57	.55	.46
GGC-cRNN	.90	.68	.60	.60	**.59**	.57
GC-cRNN	.95	.56	.56	.58	.55	.51
GGC-LOO	.60	.69	.54	.52	.54	.58

Table 1: Graph Ganger Cusality accuracy on the 5 Dream3 10-gene datasets and 1 simulated dataset. Comparable results from GC-cLSTM and GC-cRNN [20] are listed.

Model	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
GGC-cLSTM	.93	.93	.94	.89	.79

Table 2: Graph Ganger Causality cLSTM accuracy on the 5 Dream3 100-gene datasets.

distance correlation function. And secondly, the optimization step penalizes these weights if they steer too far away from these values. We found that using informative network weights, the model tends to better capture the true causality structure behind the time-series data. The deep learning architecture proposed here for studying causality in dynamic networks was able to reconstruct the Lorenz-96 simulated dataset to 99% accuracy and has achieved good results on more challenging Dream3 gene regulatory networks time-series datasets. Future work can include the use of complex models with more hidden layers to detect the large-range dependencies in the high-dimensional setting to further enhance the performance and interpretation of the Granger causality framework. To resolve computation issues in high-dimensional and large dataset settings, parallel and distributed cloud computing can be helpful.

References

[1] Dream challenges. https://dreamchallenges.org/dream-3-in-silico-network-challenge/. Accessed: 2021-08-30.

[2] Juan Camilo Castro, Ivan Valdés, Laura Natalia Gonzalez-García, Giovanna Danies, Silvia Cañas, Flavia Vischi Winck, Carlos Eduardo Ñústez, Silvia Restrepo, and Diego Mauricio Riaño-Pachón. Gene regulatory networks on transfer entropy (grnte): a novel approach to reconstruct gene regulatory interactions applied to a case study for the plant pathogen phytophthora infestans. *Theoretical Biology and Medical Modelling*, 16(1):1–15, 2019.

[3] Xing Chen and Li Huang. Lrsslmda: Laplacian regularized sparse subspace learning for mirna-disease association prediction. *PLoS computational biology*, 13(12):e1005912, 2017.

[4] Jie Dong, Mengyuan Wang, Xiong Zhang, Liang Ma, and Kaixiang Peng. Joint data-driven fault diagnosis integrating causality graph with statistical process monitoring for complex industrial processes. *IEEE Access*, 5:25217–25225, 2017.

[5] Justin D Finkle, Jia J Wu, and Neda Bagheri. Windowed granger causal inference strategy improves discovery of gene regulatory networks. *Proceedings of the National Academy of Sciences*, 115(9):2252–2257, 2018.

[6] C Granger. nvestigating causal relations by econometric models and crossspectral methods. *Econometrica*, 37:424–438, 1969.

[7] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[8] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[9] Qiang Ji, Elie Bouri, Rangan Gupta, and David Roubaud. Network causality structures among bitcoin and other financial assets: A directed acyclic graph approach. *The Quarterly Review of Economics and Finance*, 70:203–213, 2018.

[10] Alireza Karimi and Mark R Paul. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4):043105, 2010.

[11] Jonathan Lu, Bianca Dumitrascu, Ian C McDowell, Brian Jo, Alejandro Barrera, Linda K Hong, Sarah M Leichter, Timothy E Reddy, and Barbara E Engelhardt. Causal network inference from gene transcriptional time-series response to glucocorticoids. *PLoS computational biology*, 17(1):e1008223, 2021.

[12] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. Drug similarity integration through attentive multi-view graph auto-encoders. *arXiv preprint arXiv:1804.10850*, 2018.

[13] Murat Okatan, Matthew A Wilson, and Emery N Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17(9):1927–1961, 2005.

[14] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

[15] Weijie Ren, Baisong Li, and Min Han. A novel granger causality method based on hsic-lasso for revealing nonlinear relationship between multivariate time series. *Physica A: Statistical Mechanics and its Applications*, 541:123245, 2020.

[16] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

[17] David I Shuman, Benjamin Ricaud, and Pierre Vandergheynst. Vertex-frequency analysis on graphs. *Applied and Computational Harmonic Analysis*, 40(2):260–291, 2016.

[18] Elsa Siggiridou and Dimitris Kugiumtzis. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64(7):1759–1773, 2015.

[19] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[20] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality. *arXiv preprint arXiv:1802.05842*, 2018.

[21] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016.

[22] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

[23] Norbert Wiener. The theory of prediction. modern mathematics for engineers. *New York*, 165, 1956.

[24] Axel Wismüller, Adora M Dsouza, M Ali Vosoughi, and Anas Abidin. Large-scale nonlinear granger causality for inferring directed dependence from short multivariate time-series data. *Scientific reports*, 11(1):1–11, 2021.

[25] Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16(1):1–18, 2015.