Relating Feed-Forward Loop Crosstalk to Robust Information Transport Across Transcriptional Networks

Khajamoinuddin Syed [*]
lnusk@.vcu.edu

Ahmed Abdelzaher [†]
abdelzaheraf@vcu.edu

Michael Mayo [‡]
Michael.L.Mayo@usace.army.mil

Preetam Ghosh [§]
pghosh@vcu.edu

ABSTRACT

Evolved biological network topologies may resist perturbances to allow for more robust information transport across larger networks in which their network motifs may play a complex role. Although the abundance of individual motifs correlate with some metrics of biological robustness, the extent to which redundant regulatory interactions affect motif connectivity and how this connectivity affects robustness is unknown. To address this problem, we applied machine learning based regression modeling to evaluate how feed-forward loops interlinked by crosstalk altered information transport across a network in terms of packets successfully routed over networks of noisy channels via NS-2 simulation. We developed 233 topological features which distinctly account for the opportunities in which two feed-forward loops may exhibit crosstalk. Random forest regression modeling was used to infer significant features from this modest configuration space. The coefficient of determination was used as a primary performance metric to rank features within our regression models. Although only a handful of features were highly ranked, we observed that, in particular, edge connected feed-forward loops correlated substantially with an improved chance for successful information transmission.

Categories and Subject Descriptors

D.2.8 [**Machine Learning**]: Metrics—*Complexity Measures, Coefficient of Determination, Regression, Feature Ranking*

Keywords

Motif Connectivity, Transcriptional Networks, Complex Networks, Crosstalk, Edge-Connected Motif

1. INTRODUCTION

Network motifs are recurrent network structures that exhibit higher statistical significance in biological networks than in random ones. In the past, they have been implicated in the tendancy for information transport to resist noisy perturbances and successfully convey the cellular state. Past studies indicate that feed-forward loop (FFL) network motifs are important, not just in terms of abundance [11], but also in terms of certain behaviors such as response time [10]. Feed-forward loop structure (Figure 1) is intriguing because it offers two ways of regulating a protein-expressing gene (node C) via two influential paths: a direct route (A to C), or an indirect path beset by a waypoint (A to B to C). This setup may be communicationally efficacious due to the signaling modality of multiple regulatory paths to protein expression of a regulated gene. We may therefore hypothsize that higher FFL abundances will lead to better information transmission performance. One central question remains:

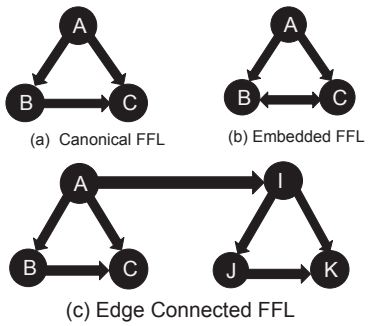

Figure 1: (Top Left) A canonical feed-forward loop is one free of additional interactions. (Top Right) Embedded feed-forward loops are contained within more complicated topological configurations. (Bottom) Feed-forward loops interrelated by sharing an edge.

[*]Corresponding author - lnusk@vcu.edu
Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA
[†]Department of Computer Science, Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA
[‡]Environmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, MS 39180
[§]Department of Computer Science, Virginia Commonwealth University, 401 W Main St, Richmond, VA, USA

BICT 2017, March 15-16, Hoboken, United States
ISBN 978-1-63190-148-5
DOI: 10.4108/eai.22-3-2017.152409

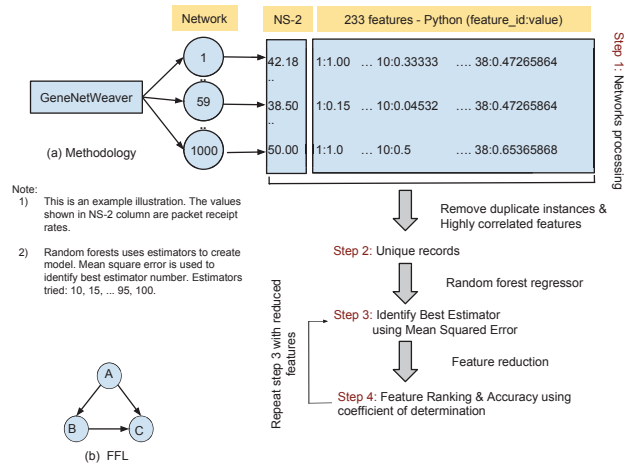

Figure 2: (a) Network extraction and NS-2 simulation methodology. (b) A feed-fowrad loop transcriptional network motif.

Do FFLs contribute signaling/communication benefits individually, synergistically in combination with others, or not at all? To address this question, we examined the extent to which feed-forward loops crosslinked by regulatory interactions (edge-connected motifs) contribute to successful information transport across biological networks, modeled as networks of noisy channels across which information packets are routed via NS-2 simulation.

Existing network robustness metrics are predominantly "static" [2, 3], in that they do not consider dynamical information transport. Chan *et al.* [2] provides an in-depth review of existing robustness metrics. Notably absent are metrics which consider motif-based features including the possibility of crosslinked feed-forward loops. Here we are concerned with the successful transmission of information packets routed across a biological network, modeled using the discrete event network simulator NS-2. These simulations account for the dynamics of information flow among the nodes in a network under controlled conditions such as channel noise and congestion-based information loss. To this effect, we define informational "robustness" as the ratio between the total number of packets received at perfectly absorbing "sink" nodes to the total number of packets emitted from potentially many source nodes. We will refer to this metric colloquially as the packet receipt rate (PRR), which accounts for network behavior resulting from graded perturbations, and is more comprehensively detailed elsewhere reports [9]. We employ discrete event simulations and machine learning techniques to develop a model trained using feature data to predict robust network topologies for information transport. We use these anlayses to rank-order the differing configurations of linked feed-forward loops, seeking to answer the following questions: Does abundance positively correlate with information-transport robustness? If so, which features are primary contributors to robustness?

2. MATERIALS AND METHODS

Our basic methodology is illustrated by Figure 2. First, subnetworks extracted (Section 2.1) from transcriptional regulatory networks of the *Escherichia coli* (*E. coli*) bacterium, and passed to the network simulator platform NS-2 (Section 2.2) to generate packet receipt rates. Next, feature values are determined using *Python*, from which we remove all the duplicate feature vectors and retain a unique feature vector with minimal PRR among all features vectors. Features are further scaled to the interval $[0, 1]$, which reduces the processing time of our regression models. However, scaling is not necessary for Random Forest regression. These data are processed (Section 2.4) into a format illustrated by Step 1 of Figure (2), upon which a random forest regression machine learning technique is applied for ranking purposes. The coefficient of detemrination is calculated to identify an optimal estimators number (Section 2.5). Before feature ranking is actually done, we perform feature selection which is a process to reduce the feature set (from the original 233 feature set). Finally, features are ranked using feature importance–a method used to determine feature significance in regression trees. Section 2.6 details the parameters used for creating random forests regression models and accuracy measurement.

2.1 Network Datasets

Directed transcriptional subnetworks from the *E. Coli* bacterium were extracted using GeneNetWeaver [13], with a total of 300, 400, and 500 total genes, and repeated 1000 times for each network size. Regulatory information was retained while disconnected network components, and autoregulatory loops were discarded. Table 1 shows the details of the network counts considered here. This step pruned the datasets down to 957, 932, and 941 networks for, respectively, the 300, 400, and 500 network sizes. This dataset is then used to explore network dynamics in two ways: a) model interactions using NS-2 (Section 2.2) and b) determine structural features of importance. Feature vectors were generated by extracting features from pruned networks, and all duplicate feature vectors were removed. The number of unique feature vectors are reported in Table 1.

2.2 NS-2 simulation setup

The problem of information flow across a biological network can be mapped onto the problem of packet transport over a wireless sensor network [4, 6, 7, 5]. In the NS-2 model, each node relays finite-sized packets of information to other nodes along outgoing edges to neighboring nodes. Packets are relayed in this manner using a flooding protocol until they reach (perfectly absorbing) sink nodes, which do not retransmit. Genes coding for transcription factor proteins, and those that do not, are represented as nodes in the network with regulatory interactions conceptualizing communication channels which determine the destination nodes of transmitted packets. Biology is inherently noisy, and we account for this by using noisy channel models wherein 10%, 20%, 35%, 50%, 60%, 75% and 90% of packets will be, on average, lost during transmission across any individual internode route. Packet receipt rate in the network is mea-

Table 1: Ttranscriptional network properties.

Size	Connected Networks	Unique Feature Vectors
300	958	163
400	933	168
500	942	157

Table 2: *F*eature Reduction from 233 features in each network

Size	Occurring Features	Uncorrelated features
300	95	50
400	98	52
500	138	57

sured as the percentage of the number of packets received at sink nodes to the number of packets sent by all source nodes. Networks with higher packet receipt rate are considered to be more *robust*. Packet receipt rates of the networks range in between 0 (least *robust*) and 100 (most *robust*).

2.3 Feature Identification

We developed toplogy-based network characteristics to understand which network qualities and features contribute primarily to information transport and routing over biological networks. While some of these characteristics, such as average shortest path, network density, and betweenness centrality have been considered before, our emphasis on using them to evaluate information transport and potential robustness of these dynamical routing processes places strong emphasis on the network dynamics. Previously, we identified fifteen different network features and ranked them using unsupervised learning techniques [8, 9]. While previous work has focused on properties of individual FFLs, we focused here on understanding how FFLs coupled by crosstalk behave within the embedding evnvironment of the network. To this effect, we developed 233 unique features (selected features in Table 3) that captures the abundance of connected FFL structures as follows. First, we identified all possible ways in which two feed-forwarded loop motifs could be connected by one or more edges; second, we counted the occurance of each pattern in the above mentioned transcriptional networks.

We used machine learning techniques to identify significant features among a list of several features, and employed different machine-learning strategies by leveraging the widely recognized *scikit* module in *Python* [12]. We do not exhaustively tabulate data on edge-connected motif abundance for

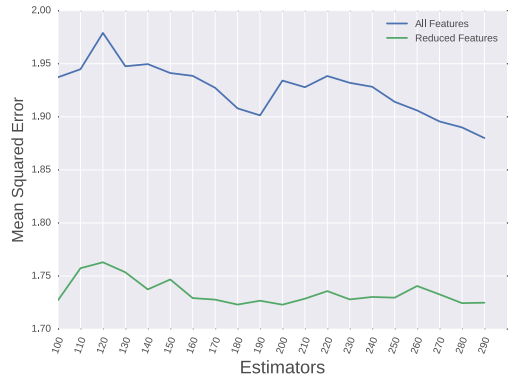

Figure 3: Mean squared error (MSE) for different number of Random Forest estimators for networks of size 300 and heavy channel loss (90%). A lower MSE here indicates a better performance.

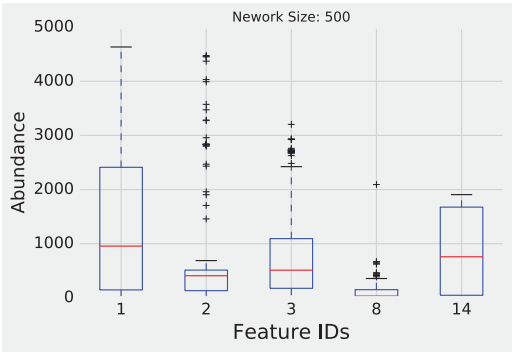

Figure 4: Feature value distribution. Refer to Table 3 for depictions of features 1, 2, 3, 8, and 14.

different network sizes here due to length considerations, but we have nevertheless provided a summary of these data in Section 6. Testing for correlations between feature abundance and feature importance is described below in Section 2.7

2.4 Data

Data is constructed similar to the procedures previously described [9]. Each network is represented as a combination of feature values, feature ids and output labels determined using NS-2 simulation. Each network (section 2.1) is represented as a combination of output labels and 233 configurations of feed-forward loops connected by crosstalk, which we term "edge-connected features." In the field of machine learning, such a combination is referred to as a feature instance. Results from NS-2 simulations are used as output labels and the corresponding features are calculated using the *networkX* [14] module in *Python*. In previous work [8, 9], we considered the problem of ranking features to be unsupervised one, and used an analysis of variance (ANOVA) F-value to determine significance of each feature. Here, however, we consider the problem to be a supervised one and retain the output labels, which range between 0 and 100, as floating points. Regression techniques are suitable when the value of output labels is continuous. Furthermore, we introduce feature selection here as an improvement from our earlier work wherein the entire feature set was used to rank features. Before creating the regression model, data is split into training and testing data in 80:20 ratio. The accuracy of regression models presented in Figure 6 is based on testing of the model created on the test data of edge connected FFL based features.

2.5 Feature Down-Selection

We selected only a subset of all 233 edge-connected features, because there is potential for some of them to be correlated with others (section 2.3) or some of them might display a higher variance. To begin, we first selected features that occurred in more than one network. The second column of Table 2 shows these feature counts. Because our aim was to deduce a minimal set of features important for information transport across these networks, we eliminated pairs of features that were positively correlated if Spearman's correlation coefficient was > 0.95. The third column of Table 2 shows these counts upon removing such correlated features. Finally, we considered different feature selection methods

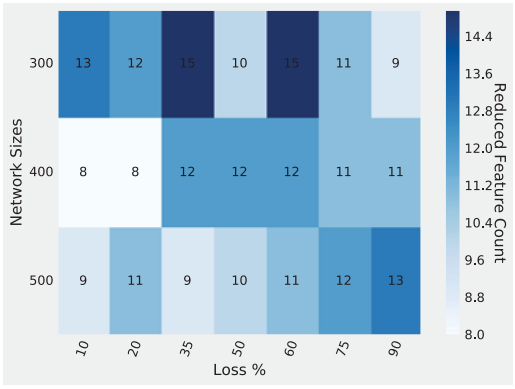

Figure 5: Selected features from a total of 233 for every model at a given network size and channel loss model, as described in 2.5

and examined those residing in the intersection. Randomized PCA was considered but ignored since it does not exploit the output label data to minimize the feature space. To this effect, a feature selection step was performed using random forests with regression.

Random forest models [1] are well-suited to solve classification and regression problems. A "random forest" refers to the trees (estimators) used by ensemble machine learning models to predict the outcome of data. Mean squared error (MSE) is used to determine the best number of estimators (number of decision trees) used in the random forests algorithm. A number of estimator (e.g, 100 to 300) incremented by steps of 10, were used here in creating the random forest model. MSE is determined for each estimator and the average of the number of estimators is used as the MSE value for that specific estimators' number. The variation in MSE noted before and after feature reduction, and shown in Figure 3 for a singular case of a 300 node network with 90% channel loss model. Before reduction, MSE is lowest for 290 estimators, whil it is lowest 200 after reduction. The estimator for which MSE is the least was selected for calculating feature importance, as shown in Figure 3.

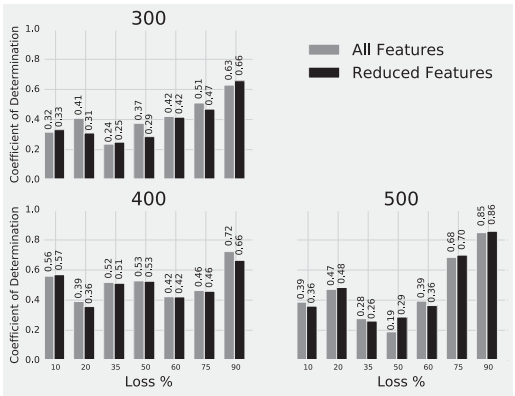

Figure 6: Coefficient of determination (COD) for edge-connected FFL features regressors model for different network sizes. A higher COD value indicates a improved performance.

Table 3: Feature abundance.

Feature ID	Symbol	Âbundance
1		41789
2		32452
3		26272
5		11068
6		10056
7		7327
8		6483
9		5398
11		4801
14		4064
16		3811
18		2994
19		2826
20		2798
22		2066
25		1784
36		996
52		458
59		422
62		357
63		348
69		281
70		271
77		210
78		210
81		194
91		152
92		152
104		113
125		60
131		51

Our analyses reveal that feature importance depends heavily on the network size and channel loss model over time. All the features with importance values ≥ 0.03 were selected to model the final regressor for prediction. Figure 5 shows the final counts of selected features for different network sizes and channel loss models.

2.6 Regression modeling

Before we carry out feature reduction, we conduct random forests regression to determine a COD calculated using all uncorrelated features identified for a given network size. Important features are selected from the set using the feature importance attribute of random forests regression. We then create a random forests regressor to predict outcomes based on the model of the new feature set, and this model is tested using the test data set.

Regressor performance is measured using the coefficient of determination (COD), which quantifies how predicted values provided by the model compare against real values. Adequate regressor models typically exhibit a COD near 1, while poorly performing models exhibit values near 0. As evident from Figure 6, the COD determined from the reduced feature set (section 2.5) either improved the model accuracy or showed no substantial difference from the set of all features. In a majority of the cases, it is evident that feature reduction did not affect performance in a negative way, suggesting that the set of reduced features plays a much stronger role in information transport in these transcriptional networks than all other features. Additionally, we observe that our models perform well at higher levels of noise or channel loss.

Figure 5 shows the number of features selected by our feature-selection process from all 233 features. The maximum number of important features was 15 for the network size 300 and channel loss model of 35% and 60%, with 8 as the least number of important features for 400 node networks operating under a channel loss model of 10% and 20%. We find that many scenarios exhibit 11 and 12 important features.

Feature important (section 2.5) is shown in Figure 7. Heat maps were generated for all the networks at channel loss models of 10%, 20%, 35%, 50%, 60%, 75%, and 90%. Figure 7 represents one such case for a network size of 500 and all loss models created with every reduced feature set. We observe that features with IDs 1, 2, 3, 8, 9, 11, 14 and 52 are important for all levels of packet loss. Additionally, features 62 and 81 are important for 75% and 90% packet loss. Topologies of these features have been collected into Table 3. Here, the abundance of each feature is provided for the largest connected component of the entire *E. coli* transcriptional network.

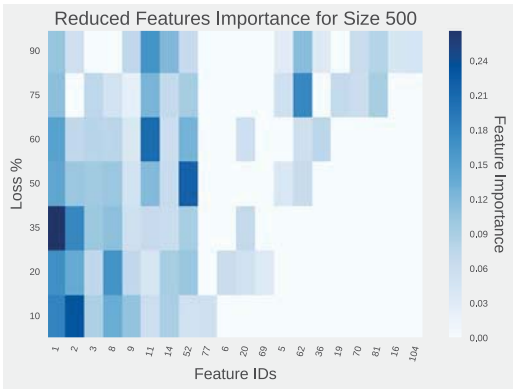

Figure 7: (a) Feature significance in size 500 networks for all loss models and reduced feature sets. The darker the color the higher the feature significance. Additionally, numbers are included to indicate feature rank; higher is better.

2.7 Feature Importance correlation with feature abundance

To test the hypothesis that high feature values correlate positively with high feature importance, we performed following task executed at network level. That is, for each network size, the significant features were identified for all models for different levels of packet loss.

1. Identify the top five features using random forest regression (feature importance as a metric);
2. Calculate the number of times each features occurs within the top five ranks at different channel loss models and network size;
3. Plot the distribution of these feature (Figure 4).

We found that correlation between abundant feature values with high variance and its importance. From all the models, features 1, 2, 3, 8, 14 are consistently in top five features, these features are strong indicators of robustness. Figure 4 shows the feature value(abundance) distribution of top five features as mentioned earlier. We can see that all the features have high abundance with high variance. It is important to note that certain features such as 62, 81, 125 make their impact distinctively in specific network sizes at specific loss scenarios. This can be attributed to the fact that these specific features might be expressed more during the network extraction step (Section 2.1). Figure 4 illustrates a boxplot of the distribution of feature values of the top five features, with outliers in the dataset marked with $+$. Feature-value distributions for other networks are not shown (see section 4).

3. DISCUSSION & CONCLUSIONS

In this paper we studied how differing topological configurations of FFL crosstalk affected the information transport success in transcriptional subnetworks of the *Escherichia coli* bacterium. We evaluated information transport according to packet transport and routing events enabled by NS-2 simulations. Random forest based regression models revealed that a handful of edge-connected FFL configurations, such as 1, 2, 3, 8 and 14 (Table 3), may have an important role in enabling the robust communication of molecular information across the subcellular transcriptional-regulatory machinery of the cell. Certain crosstalk configurations appeared differentially important under varying noise levels inherent to the communication channels. Understanding how noise interferes with communicating the cellular state to distal molecular processes is a great challenge, because the cell is a dynamically evolving environment that continually produces and destroys molecular components from which signaling success is not guaranteed.

Extensions of this work involve investigations in larger *E. coli* transcriptional subnetworks, to explore whether or not trends in feature significance scales with increasing network complexity. Furthermore, we intend to extend our anlyses to the transcriptiona-regulatory networks of the baker's yeast *Saccharomyces cerevisiae*. Previous results [9] reveal that feature significance varies from one organism to another and scales across network size and perturbation conditions. As we fine-tune our regression models it is also important to focus on moderately sized networks (e.g., 300 and 500 nodes) with larger channel loss models (e.g., 35% and 50%), to better understand why our regression models did not adequately perform.

Finally, the present work will provide a foundation for the biological network community to better understand the functional role of crosstalk between smaller transcriptional network motifs. In addition, the engineering community may benefit from knowledge that certain network topologies provide more robust communication platforms, transforming the difficult problem of information-preserving dynamical routing across terrain and environmental obstacles into one concerned only with short-range topological interactions.

4. ADDITIONAL MATERIAL

Datasets are available for research purposes at:

http://github.com/syedkm/EdgeConnectedMotifs.

In addition, this address provides results for all the channel loss models not presented here due to space considerations. Sensitivity analyses for variation in mean square error, mean absolute error, and explained variance are also provided.

5. ACKNOWLEDGMENTS

Funding was provided by the US Army's Environmental Quality and Installations 6.1 Basic Research program. Opinions, interpretations, conclusions, and recommendations are those of the author(s) and are not necessarily endorsed by the US Army.

6. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] H. Chan, L. Akoglu, and H. Tong. Make it or break it: manipulating robustness in large networks. In *Proceedings of the 2014 SIAM Data Mining Conference*, pages 325–333. SIAM, 2014.

[3] J. A. de la Peña, I. Gutman, and J. Rada. Estimating the estrada index. *Linear Algebra and its Applications*, 427(1):70–76, 2007.

[4] P. Ghosh, M. Mayo, V. Chaitankar, T. Habib, E. Perkins, and S. K. Das. Principles of genomic robustness inspire fault-tolerant wsn topologies: a network science based case study. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, pages 160–165. IEEE, 2011.

[5] S. Ghosh, P. Ghosh, K. Basu, and S. K. Das. Gama : An evolutionary algorithmic approach for the design of mesh-based radio access networks. In *Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on*, pages 374–381. IEEE, 2005.

[6] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. Perkins, and S. K. Das. Performance of wireless sensor topologies inspired by e. coli genetic networks. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 302–307. IEEE, 2012.

[7] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. J. Perkins, and S. K. Das. Leveraging the robustness of genetic networks: a case study on bio-inspired wireless sensor network topologies. *Journal of Ambient Intelligence and Humanized Computing*, 5(3):323–339, 2014.

[8] B. K. Kamapantula, M. Mayo, E. Perkins, A. F. Abdelzaher, and P. Ghosh. Feature ranking in transcriptional networks: Packet receipt as a dynamical metric. In *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, BICT '14, pages 1–8, 2014.

[9] B. K. Kamapantula, M. Mayo, E. Perkins, and P. Ghosh. Dynamical impacts from structural redundancy of transcriptional motifs in gene-regulatory networks. In *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, BICT '14, pages 199–206, 2014.

[10] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[11] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] T. Schaffter, D. Marbach, and D. Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[14] D. A. Schult and P. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.