**The FFL attachment algorithm for generating realistic gene regulatory networks**

Erik K. Zhivkoplias1**⍏**, Oleg Vavulov2**⍏**, Thomas Hillerton1, Erik L. L. Sonnhammer1\*

1Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Box 1031, 17121 Solna, Sweden

2Bioinformatics Institute, St. Petersburg, Russia

**⍏**These authors contributed equally

\*To whom correspondence should be addressed.

**Abstract**

Gene-gene and gene-protein hidden regulatory relationships form a gene regulatory network (GRN) that controls the cellular response to changes in the environment. A number of inference methods to reverse engineer the original network from large-scale expression data has recently been developed. However, the absence of ground-truth networks when evaluating the performance makes realistic simulations necessary. One aspect of this is that local network motif analysis of real GRNs indicates that the feed-forward loop (FFL), is significantly enriched. To simulate this properly, we developed a novel motif-based preferential attachment algorithm, FFLatt, which outperformed the popular GeneNetWeaver network generation tool in reproducing the FFL motif occurrence observed in literature-based biological networks. It also preserves important topological properties such as scale-free distribution, sparsity, and average in/out-degree per node. We conclude that FFLatt is well-suited as a network generation module of a data generation framework with the aim to serve for fair and robust performance evaluation of network inference methods.

Keywords: network biology, gene regulatory networks, gene-gene interaction, network motif structure, network generation, network simulation, benchmarking

Availability: <https://bitbucket.org/sonnhammergrni/fflatt/>

**Introduction**

Understanding large-scale biological relationships between genes and the proteins they encode remains a great challenge in systems biology. The wide availability of system-level expression datasets gave rise to a variety of reverse-engineering methods that aim to reconstruct the hidden regulatory gene–gene and gene–protein relationships. Such relationships form a gene regulatory network (GRN) that regulates developmental processes in organisms and controls adaptation to changes in the environment [(Davidson 2010)](https://paperpile.com/c/1g2hAN/Ynze). By contrast with other networks in biological systems, GRNs are harder to validate as the interactions that occur between genes usually involve indirect interactions through biological molecules making the interaction hard to detect and quantify. The incompleteness and scarcity of ground-truth networks results in problems when evaluating the performance of methods that seek to infer GRNs from large-scale expression data [(Emmert-Streib and Dehmer 2018)](https://paperpile.com/c/1g2hAN/d7m8).

The problem of inferring a gene regulatory network from gene expression data has received significant attention. A great variety of GRN inference methods have been introduced [(Margolin et al. 2006; Faith et al. 2007; Huynh-Thu et al. 2010; Friedman, Hastie, and Tibshirani 2010; Zavlanos et al. 2011)](https://paperpile.com/c/1g2hAN/y84T+zmu6+6y8v+3Ra7+stP0) to tackle this problem. It was also the focus of four separate Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges, with DREAM5 being the most recent one [(Marbach et al. 2012)](https://paperpile.com/c/1g2hAN/4jdq). Newer, more advanced algorithms require not only expression data but also utilize additional information such as experimentally validated interactions and Gene Ontology terms [(Chouvardas, Kollias, and Nikolaou 2016)](https://paperpile.com/c/1g2hAN/mOQk), structures of genomic datasets and network topology [(Siahpirani and Roy 2017)](https://paperpile.com/c/1g2hAN/eLIT) or Bayesian model that incorporates multiple evidence sources [(Young, Raftery, and Yeung 2014)](https://paperpile.com/c/1g2hAN/ZXRv). Despite this, for most methods the performance on real experimental datasets remains modest [(Marbach et al. 2012; Chen and Mar 2018; Pratapa et al., n.d.)](https://paperpile.com/c/1g2hAN/4jdq+3foL+zcTV).

Regardless of the method used, it is important to fairly assess its performance with respect to other methods. As some methods can only predict Boolean networks, assessment should be done in terms of binary error classification such as the number of false positives and false negatives. In addition to this, the information about transcriptional interactions is usually only available in the binary form. Boolean networks can only be defined by their topology, which is why it makes it essential to understand the structure of GRN graphs. It’s also worth pointing out that most GRN inference methods could only predict the static network structure, which implies that *in-silico* generated data should also possess biological stability.

While the true Boolean structure of GRN is usually not known, all GRNs still share some known structural topology: the scale-free property [(Barabasi and Albert 1999)](https://paperpile.com/c/1g2hAN/z8ZP), where the average node degree follow a power-law degree distribution, and/or small world property [(Watts and Strogatz 1998)](https://paperpile.com/c/1g2hAN/B7wJ), where nodes form distinct clusters in which they are connected to each other in lattice rings. These properties are different from random graphs where node degree is uniformly distributed across all nodes in the system. Some attempts to model GRNs have been made by implementing methods that generate random [(Watts and Strogatz 1998; Mendes, Sha, and Ye 2003)](https://paperpile.com/c/1g2hAN/B7wJ+nK6W) or scale-free artificial gene networks with given sets of parameters but eventually methods that are based on the idea of subnetwork-selection gained more popularity [(Van den Bulcke et al. 2006; Schaffter, Marbach, and Floreano 2011)](https://paperpile.com/c/1g2hAN/STjv+PBVb). The most popular method, GeneNetWeaver(GNW, [(Schaffter, Marbach, and Floreano 2011)](https://paperpile.com/c/1g2hAN/PBVb))was used to generate *in silico* networks for DREAM challenges. GNW allows for knock-out and knock-down perturbation designs when generating networks and expression data, and control of the number of nodes, including the number of TFs, based on a user-defined input network. To generate expression data it utilizes a non-linear ordinary differential equations (ODE) model for gene expression and stochastic differential equations (SDEs) for molecular noise generation with Hill function-based kinetics.

Global network dynamics is shaped by gene regulatory patterns that are more frequent in GRNs than in other networks [(Shen-Orr et al. 2002; Milo 2002)](https://paperpile.com/c/1g2hAN/2eRV+qxW1) and may carry information-processing functions. These local patterns, or motifs, do not result in emergence of specific patterns in gene expression but rather determine dynamical boundaries of the phase space of the system [(Ahnert and Fink 2016)](https://paperpile.com/c/1g2hAN/k7Bw). It was suggested that some motifs could be particularly important for network dynamics and so they become overrepresented and drive the evolution of the networks [(Prill, Iglesias, and Levchenko 2005)](https://paperpile.com/c/1g2hAN/dgp1). Due to this, simulating the network structure that preserves the overrepresentation of the motifs is of utmost importance for capturing realistic dynamics of GRNs. The idea of building gene regulatory networks by using motifs as building blocks was first introduced by [(Abdelzaher et al. 2015)](https://paperpile.com/c/1g2hAN/zGmk) after they observed that most of the edges in the *E. coli* network were involved in downlink motif, and so it was hypothesized that it could have a particular importance for the evolution of the GRN topology in *E. coli* [*(Abdelzaher et al. 2015)*](https://paperpile.com/c/1g2hAN/zGmk).

In the present study the significance of 3-node motifs in four directed GRNs based on experimentally verified transcriptional interaction databases were evaluated. In agreement with previous studies [(Lee et al. 2002; Brivanlou 2005; Milo 2002)](https://paperpile.com/c/1g2hAN/dqjm+D48O+qxW1), it was found that the feed-forward loop (FFL) motif is the only one that is overrepresented. We therefore developed a novel motif-based preferential attachment algorithm for simulating realistic structures of GRNs, structures that are enriched with the FFL motif. We finally conclude that FFL-based *in silico* networks demonstrate structural properties that are common for biological gene regulatory networks, and suggest that they can be used for fair and robust evaluation of a wide variety of GRN inference algorithms performance.

**Methods**

### **Transcriptional interaction databases**

Three biological databases that contain information of experimentally validated transcriptional regulation were chosen as ground-truth networks: RegulonDB [(Santos-Zavaleta et al. 2019)](https://paperpile.com/c/1g2hAN/kB9t) for *E. coli*, [(Balaji et al. 2006)](https://paperpile.com/c/1g2hAN/AbYR) for *S. cerevisiae*, and TRRUST v2 [(Han et al. 2018)](https://paperpile.com/c/1g2hAN/note) for *M. musculus* and *H. sapiens* TF-gene target regulatory relationships.

### **Motif-node participation and Motif enrichment**

We chose to test for node-motif participation for all possible connected three-node motifs with no reciprocal links between them (Figure 1). To calculate the motif-node counts, *Nreal*, for every node in the network we calculated the presence of a given node in all different roles of a given motif, *N(i)*. and so for a set of nodes {*1=1, …, M}* in the network of size *M* it could be framed as:

For example, node *a* could either participate in Role 1 (2 outgoing edges, 0 incoming), Role 2 (1 outgoing edge, 1 incoming), Role 3 (0 outgoing edges, 2 incoming) of FFL motif 1 but at the same time participate in different role of other FFL motif 2 (Figure 2).

To test for motif enrichment, we calculated Z-score for every motif type:

where *Nreal* is the number of motif counts in the original network, *μshuffled* and *σshuffled* is the mean and standard deviation of motif counts in the distribution of shuffled networks. Every network was shuffled with a persevered in/out-degree for all nodes until at least 80% of edges in the original network were swapped. To calculate the mean and standard deviation of motif counts in the shuffled networks every network was shuffled 10 000 times.

### **Algorithm description**

The FFL-based generation algorithm starts with nucleation where the target, FFL-enriched network, is first searched for FFL motifs. FLL-based subnetwork of predefined size (a default minimum size equal to 20 nodes) with all FFL connected via shared nodes is then used as a substrate when attaching new edges and nodes to the growing network. The outline of the algorithm is present graphically and in the form of pseudocode (Figure 3, Figure 4).

Once the substrate is selected the algorithm adds nodes and edges iteratively such that at every iteration, a candidate node is selected with a random uniform probability. Once selected, one of the four attachment rules (R1, R2, R3, R4) is applied (Figure 4) based on four predetermined probability scores (*p1, p2, p3, p4*) that add up to 1. The iterations are repeated until the required number of nodes in the network is reached.

If the random float number *r1 (RandomUniform[0,1])* is less or equal to *p1* then R1 is picked. For the R1 rule we applied the modified preferential attachment algorithm from Mayo et al. (2012) with power-law kernel:

where *Ki* denotes node-degree connectivity, *∏* is the probability that a new node will be connected to node *i*, and *ɣ* is a parameter that controls the out-degree distribution.

If *r1* is greater than *p1* then one of the motif-based preferential attachment rules (R2, R3 or R4) is applied, and so *1-p1* corresponds to the desired percentage of nodes that participate in FFL motifs. For R2-R4 rules, one of the already existing FFL-motifs is picked based on it’s connectivity with the others.

Once the candidate motif and rule are chosen, a new random float number, *r2 (RandomUniform[0,1])*, is generated. If 0 < *r2* ≼ *p2,* R2 rule is applied. In that case, two new edges and one new node will be added to the existing node so the new FFL motif is formed. If *r2 > p2*, either one of the two R3 or R4 rules is selected with the equal probability being chosen. For R3 rule, one new outgoing edge and one new incoming edge are added to the candidate node. For R4 rule, only one outgoing edge is added to the candidate node. See Fig. 4 for details.

All nodes have to have out-degree connectivity [](https://www.codecogs.com/eqnedit.php?latex=\mathcal{L}_{out}" \l "0) smaller than threshold, *Kmax*, after which no new outgoing edges can be added. If the candidate motif doesn’t satisfy the conditions for a chosen FFL-attachment rule, the search for a candidate motif is repeated until it meets the rule conditions. If a new motif is created, the library with FFL-motifs is updated.

When the desired network size is reached, the algorithm controls for sparsity (average number of connections per gene) until it reaches the set sparsity level. If the network is too dense, edges are removed based on out-degree node connectivity so the edge is more likely to be removed if it is attached to a node with a high out-degree. If the network is too sparse, outgoing edges are added based on the out-degree connectivity and connected nodes are selected randomly.

### **Network generation**

For network simulation comparison five algorithms were chosen: FFLatt (developed in present study), GeneNetWeaver (GNW; [(Schaffter, Marbach, and Floreano 2011)](https://paperpile.com/c/1g2hAN/PBVb)), the modified NetworkX directed scale-free graph algorithm (NetworkX graph; Bollobás et al. 2003) so it could control for network sparsity, and sparse uniformly distributed random binary matrix with and without allowing for feedback loops in the network (DAG and RandG; [(Guo and Amir 2021b)](https://paperpile.com/c/1g2hAN/afjs)). For network generation of different sizes with FFLatt, the set of transcriptional interaction graph properties estimated from the *E.coli* transcriptional interaction network (Table2) was used. For network generation of different sizes with other algorithms (except GNW), only network size and sparsity parameters were taken into account as only controllable parameters available. For network generation/subselection with GNW the following (default) parameters were used: *--random-seed*, *--greedy-selection,* *--keep-self-interactions* as well as the size of the subtracted network.

When mimicking the *E.coli* transcription network model, all three-node cycles were disrupted, by removal of one edge, as they are absent in the target network. The removal was done by deleting the outgoing edge of the node with the highest out-degree and an edge was instead attached to a random node with a probability based on the connectivity of each node.

For stability analysis, self-loops (if any) were removed from network graphs generated with above mentioned algorithms before applying the stability analysis model.

### **Stability analysis model**

To measure the stability of networks, we utilized the model developed by [(Guo and Amir 2021a)](https://paperpile.com/c/1g2hAN/nEXE) that explores how the dynamics of protein and mRNA concentrations control the transcriptional regulation. The model allows for multiple proteins acting on the same gene, and is defined by the authors as:

where *gi* and gi0 is the effective gene copy number of gene *i with and without input of other genes respectively*, *cj* is the concentration of transcription factor *j*, and *aij* relates to the strength of the regulation of gene *i*by *cj*. The functional relationship between the transcription factor and target gene, *fij*, is modelled as a sigmoid Hill function:

where *h* is the saturation binding coefficient (*h*>0), and *K* is the expression threshold.

We assume that the mRNA degradation rate is much faster than that of proteins, as that was suggested by [(Guo and Amir 2021a)](https://paperpile.com/c/1g2hAN/nEXE), and therefore followed with their conclusion that as the dynamics of mRNA concentration ≈ 0, the dynamics of transcription factors concentrations can be simplified as:

## 

where *φ is the fraction of RNA polymerases that binds gene i, kp* is the transcription rate of ribosomes, and *cr* is the ribosomal concentration. In such case, the stability of the transcriptional regulatory network is dependent on the Jacobian matrix A of size NxN:

whereis the steady-state ribosomal concentration, *M* is the interaction matrix, *I* is the identity matrix, and N is the number of genes in the system. The system is stable if the maximal real part of all eigenvalues of M, *λM*, is smaller than 1, i.e. all eigenvalues of A are negative.

We applied the Guo and Amir model to all network graphs simulated with different algorithms. Each graph, in a form of adjacency matrix, was supplied as A matrix. For each replicate of a different size generated with a given algorithm, we repeated signing the network graph with respective link strengths 10 times. To focus on the effect of the GRN structure and FFL content on stability, we forced the distribution of link strengths of all GRNs to be similar. This was done by randomly setting half of the links to be upregulated and the other half downregulated (setting *max(aij)* and *min(aij)* to 1.5 and -1.5 respectively as boundaries of a normal distribution). In every trial, we were first calculated the ribosomal concentrations with which the system reaches its steady state. Once the steady state was found, λM was calculated, the highest eigenvalue of λM was stored and compared across networks of different sizes.

## **Results**

### **FFL is the only enriched three-node motif in transcriptional interaction databases**

Of all possible 3-gene network motifs with 2 or 3 unidirectional links, we found a strong enrichment of the FFL motif in the networks studied here, which are networks that capture direct TF-TG interactions (Table S1). This was previously shown for *E. coli* [(Milo 2002)](https://paperpile.com/c/1g2hAN/qxW1) and *S. cerevisiae* [(Lee et al. 2002)](https://paperpile.com/c/1g2hAN/dqjm). Interestingly, we also found that the cascade, uplink, and downlink motifs were consistently and significantly (P-value<0.05) depleted in all four target networks. All depleted motifs are a subset of a non-directed 3-node motif with two edges, where one node is connected to two other ones. A 3-node motif with two edges was already shown to be significantly depleted in other biological networks, for instance in a protein structure network and a human brain functional network [(Mirzasoleiman and Jalili 2011)](https://paperpile.com/c/1g2hAN/ybbW). However, how the depletion of the motif might potentially contribute to the function of the gene circuitry, and how it relates to the evolution of gene regulatory networks remains to be answered.

We found that FFL is the only motif with a positive Z-score, meaning it is enriched. We calculated how many of FFL-motifs share common nodes with each other, and found that this number ranges from 99.1 % in the *E. coli* GRN to 100% in *S. cerevisiae* (Table 1). After that, we calculated the fraction of nodes that participate in FFL motifs and found out that it ranges from 27% to 37.4%. This inspired us to develop a GRN generation algorithm that attaches nodes that form FFL motifs to this degree. For each GRN we also calculated the average edges per node, here referred to as sparsity, and in- and out-degree, these values were used as targets when mimicking GRNs.

Each regulatory interaction in the FFL motif could be either positive or negative, i.e. activating or inhibiting, resulting in eight different types that can act as e.g. accelerators, delay-generators or pulsers [(Mangan and Alon 2003)](https://paperpile.com/c/1g2hAN/siLg), resulting in different dynamics of gene circuits. Given the wide variety of FFL types and their importance to GRN dynamics, an unsigned *in silico* GRN graph needs a large number of FFLs to accommodate these. A combination of the eight signed types of FFL motifs will in turn reflect a realistic flow of GRN circuits.

We generated a set of GRNs of different sizes in a range from 500 to 1500 nodes, 10 replicates for each size, using five different algorithms: FFLatt , GNW [(Schaffter, Marbach, and Floreano 2011)](https://paperpile.com/c/1g2hAN/PBVb), NetworkX graph RandG, and DAG. For each algorithm we estimated four properties: the number of nodes that participate in FFL-motifs, network sparsity, average in- and out- degree within the network. We repeated these simulations for all organisms, as all four transcriptional databases have different graph properties. The results of our simulations for the *E.coli* graph are shown in Figure 6, and for other organisms in Supplementary materials (Fig. S2, Fig. S3, Fig. S4).

To assess the performance accuracy of network inference algorithms, the topological parameters such as in- and out-degree distribution and sparsity should be controlled when simulating data for benchmark analysis. We found that sparsity as well as out-degree of artificial networks generated with the subnetwork-selection based GNW algorithm deviates from set parameters of target networks, when we tried to mimic topological properties of *E. coli* GRN (Fig. 6B,6D) in networks of sizes 500 and 750, of *S. cerevisiae* (Fig. S1B, S1D) in networks of sizes 500, and in networks of all sizes of *M. musculus* and *H. sapiens* GRNs (Fig. S2B,S2D and S3B, S3D). While this alone does not indicate a poor performance of GNW algorithm, it does advocate for the necessity of network generation algorithms with a controlled set of topological parameters.

More importantly, despite subsetting networks from true GRNs with GNW algorithm, we found a significant underrepresentation of FFL motif-node participation counts in artificial networks of sizes 500, 750, and 1000 when mimicking *E.coli* GRN (Fig. 5A) in comparison with FFLatt networks. The similar results were obtained when mimicking GRNs of other organisms (Fig S1A,S2A,S3A). To confirm our findings, we performed motif enrichment analysis on simulated networks as we did with true GRNs (Fig 7). We found that the Z-score for FFL motifs in networks generated with the GNW algorithm ranges from -1.5 to 0.4. It suggests that FFL-motifs are not overrepresented in GNW networks, differing from what was shown for the true target networks. In networks generated with other algorithms FFLs were also not overrepresented. In contrast, the Z-score of the FFL motif in FFLatt networks ranges from 2.95 to 4.98.

We also noted that the FFLatt networks were enriched with three other motifs: uplinks, downlinks and cascades whereas in GNW networks they are usually depleted as well as in true GRNs. To study if this overrepresentation in network topological structure might affect the network dynamics we decided to do the stability analysis test, as described in Methods.

### **Topology and motif composition affect network stability**.

In biology, random matrix theory (RMT) is known from R. May's research on the stability of large biological systems [(May 1972)](https://paperpile.com/c/1g2hAN/4wU4) (May 1972). He demonstrated that the stability of a large ecological system have to satisfy the following inequality:

where α is the self-regulation term (equivalent to carrying capacity), σ is the interaction strength, and *C* is the density of interactions. For every off-diagonal element of the matrix it’s *1-C* probability to be equal to zero, and *C* probability to be drawn from random distribution with mean=0 and variance=σ2. Therefore, the larger a system gets the more unstable it becomes if the sparsity and interaction strength is not scaled down accordingly. May’s approach has been proven to be highly valuable to other biological networks (cite some general biophysics), including those that aim to describe gene regulations [(Prill, Iglesias, and Levchenko 2005; Stone 2018)](https://paperpile.com/c/1g2hAN/dgp1+4C63).

It was earlier suggested that motif composition contributes to fault-tolerance in transcriptional networks (Roy et al., 2020). To test if the structural composition is important for stability in artificially generated networks, we analysed the stability of the network models as they increased in size. As expected, all GRNs with fixed sparsity and uniform interaction strength became more fragile when increasing in size. We found that GRNs with different motif profiles demonstrated different levels of network stability (Fig. 8). The *randG* GRNs that were neither enriched nor depleted with any 3-node motifs (Fig. 7) were far less stable than the other ones. We found that on average NetworkX and FFLatt GRNs demonstrated better steady-state stability compared to DAG GRNs. We note that NetworkX, GNW and FFLatt GRNs have different network motif abundances, such as depleted and enriched FFL motifs respectively, yet they show similar stability. The abundance of the FFL motif therefore does not seem to be a major factor for network stability, which is congruent with previous findings about non importance of the FFL motif to system robustness under random node failure test [(Abdelzaher et al. 2015)](https://paperpile.com/c/1g2hAN/zGmk).

We also noticed that the slope of two lines that represent size-dependent stability of DAG and RandG GRNs is different from others. Ass all networks increase in size, DAG and RandG networks are more likely to become less stable than others. We calculated the degree-distribution of the same sets of artificial network graphs that we generated. Since RandG and DAG networks are sparse uniformly distributed random binary matrices, the degree-distribution doesn’t follow the power law and so the networks are not scale-free (Fig. 9). This suggests that a scale-free topology which has been previously found to be central for creating a robust system, protecting the GRN from random mutations [(Greenbury et al. 2010)](https://paperpile.com/c/1g2hAN/a0Xa) can in fact help to stabilize gene regulatory systems so it reaches a viable state after perturbation.

## **Discussion**

We show that the motif profile and topological properties of FFLatt network graphs demonstrate the biological stability comparable with other models, such as NetworkX and GNW algorithms. It is particularly important for network inference methods working with steady-state gene expression data as many of them, such as Least-Squares with Cut-Off (LSCO; [(Tjärnberg et al. 2013)](https://paperpile.com/c/1g2hAN/QFDx), LASSO [(Tibshirani 1996)](https://paperpile.com/c/1g2hAN/Kn5C); [(Friedman, Hastie, and Tibshirani 2010)](https://paperpile.com/c/1g2hAN/3Ra7), LASSO-VAR [(Larvie et al. 2016)](https://paperpile.com/c/1g2hAN/YjbK), and GENIE3 [(Huynh-Thu et al. 2010)](https://paperpile.com/c/1g2hAN/6y8v) aim to infer a stable static network from steady-state data. To summarize, the FFLatt graph generation algorithm provides an opportunity to simulate biologically meaningful network graphs that could be wired with realistic biological dynamics.

We couldn’t find evidence that different three-node motif profiles affect network stability. NetworkX, GNW, and FFLatt profiles are fairly different yet they demonstrated comparable stability across different sizes. While being out of scope of our study, it remains an interesting question how composition of more complex and higher-order structures known to be present in GRNs [(Benson, Gleich, and Leskovec 2016)](https://paperpile.com/c/1g2hAN/yp6r); [(Gorochowski, Grierson, and di Bernardo 2018)](https://paperpile.com/c/1g2hAN/AQkb) could contribute to stability of the system.

In this article we only focus on the proof of concept of the FFL attachment algorithm to demonstrate its necessity and feasibility. However, to increase model performance, it could be extended with other parameters. For example, to better capture “small world” [(Watts and Strogatz 2011)](https://paperpile.com/c/1g2hAN/nyO2) structural properties that are known to be present in biological networks one of such parameters could be a desired number of biological modules so that within each graph connectivity is higher than in between. The clustering algorithm should, however, be biologically motivated so the connection between modular graph structure and expression dynamics is clear.

Despite a continued uncertainty of how structural properties and functional modularity of GRNs relate to each other, some patterns such as FFLs are known to be key signatures of transcriptional regulation networks. ere we developed a novel, time-efficient algorithm that generates biologically realistic structures of large artificial gene regulatory networks with controlled size, sparsity, topology, and number of FFLs. FFLatt graphs are unsigned and so, they can exhibit a wide range of dynamical structures. They could be used as input to already established tools based on Hill function kinetics (such as GeneNetWeaver). Potentially, it could also become a part of future deep learning frameworks that aim to model gene expression from DNA sequence [(Zrimec et al. 2020)](https://paperpile.com/c/1g2hAN/vzVx); [(Avsec et al. 2021)](https://paperpile.com/c/1g2hAN/zwGY). In such frameworks, FFLatt networks could be used as a deep learning model constraint to incorporate prior knowledge of each node participation in FFL motifs. As a result, we hope to contribute to the development of benchmarking tools that could fairly and accurately evaluate the performance of GRN inference methods.

## **Author Contributions**

EK-Z and O-V devised and implemented the algorithm. EK-Z and T-H performed the calculations, analyzed the results, contributed to the discussion, designed the figures and wrote the manuscript. ELL-S participated in the design and coordination of the study, contributed to the discussion and design of figures, supervised and reviewed the writing of the manuscript. All authors read and approved the final version of the manuscript.

## **Conflict of Interest Statement**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## **Acknowledgments**

We thank the Swedish Strategic Research Foundation for financial support.

**References**

[Abdelzaher, Ahmed F., Michael L. Mayo, Edward J. Perkins, and Preetam Ghosh. 2015. “Contribution of Canonical Feed-Forward Loop Motifs on the Fault-Tolerance and Information Transport Efficiency of Transcriptional Regulatory Networks.” *Nano Communication Networks*. https://doi.org/](http://paperpile.com/b/1g2hAN/zGmk)[10.1016/j.nancom.2015.04.002](http://dx.doi.org/10.1016/j.nancom.2015.04.002)[.](http://paperpile.com/b/1g2hAN/zGmk)

[Ahnert, S. E., and T. M. A. Fink. 2016. “Form and Function in Gene Regulatory Networks: The Structure of Network Motifs Determines Fundamental Properties of Their Dynamical State Space.” *Journal of the Royal Society, Interface / the Royal Society* 13 (120). https://doi.org/](http://paperpile.com/b/1g2hAN/k7Bw)[10.1098/rsif.2016.0179](http://dx.doi.org/10.1098/rsif.2016.0179)[.](http://paperpile.com/b/1g2hAN/k7Bw)

[Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. “Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions.” *Nature Methods* 18 (10): 1196–1203.](http://paperpile.com/b/1g2hAN/zwGY)

[Balaji, S., M. Madan Babu, Lakshminarayan M. Iyer, Nicholas M. Luscombe, and L. Aravind. 2006. “Comprehensive Analysis of Combinatorial Regulation Using the Transcriptional Regulatory Network of Yeast.” *Journal of Molecular Biology*. https://doi.org/](http://paperpile.com/b/1g2hAN/AbYR)[10.1016/j.jmb.2006.04.029](http://dx.doi.org/10.1016/j.jmb.2006.04.029)[.](http://paperpile.com/b/1g2hAN/AbYR)

[Barabasi, A. L., and R. Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286 (5439): 509–12.](http://paperpile.com/b/1g2hAN/z8ZP)

[Benson, Austin R., David F. Gleich, and Jure Leskovec. 2016. “Higher-Order Organization of Complex Networks.” *Science* 353 (6295): 163–66.](http://paperpile.com/b/1g2hAN/yp6r)

[Brivanlou, Ali H. 2005. “Faculty Opinions Recommendation of Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells.” *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*. https://doi.org/](http://paperpile.com/b/1g2hAN/D48O)[10.3410/f.1027950.334843](http://dx.doi.org/10.3410/f.1027950.334843)[.](http://paperpile.com/b/1g2hAN/D48O)

[Chen, Shuonan, and Jessica C. Mar. 2018. “Evaluating Methods of Inferring Gene Regulatory Networks Highlights Their Lack of Performance for Single Cell Gene Expression Data.” *BMC Bioinformatics* 19 (1): 232.](http://paperpile.com/b/1g2hAN/3foL)

[Chouvardas, Panagiotis, George Kollias, and Christoforos Nikolaou. 2016. “Inferring Active Regulatory Networks from Gene Expression Data Using a Combination of Prior Knowledge and Enrichment Analysis.” *BMC Bioinformatics* 17 Suppl 5 (June): 181.](http://paperpile.com/b/1g2hAN/mOQk)

[Davidson, Eric H. 2010. “Emerging Properties of Animal Gene Regulatory Networks.” *Nature* 468 (7326): 911–20.](http://paperpile.com/b/1g2hAN/Ynze)

[Emmert-Streib, Frank, and Matthias Dehmer. 2018. “Inference of Genome-Scale Gene Regulatory Networks: Are There Differences in Biological and Clinical Validations?” *Machine Learning and Knowledge Extraction*. https://doi.org/](http://paperpile.com/b/1g2hAN/d7m8)[10.3390/make1010008](http://dx.doi.org/10.3390/make1010008)[.](http://paperpile.com/b/1g2hAN/d7m8)

[Faith, Jeremiah J., Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. 2007. “Large-Scale Mapping and Validation of Escherichia Coli Transcriptional Regulation from a Compendium of Expression Profiles.” *PLoS Biology* 5 (1): e8.](http://paperpile.com/b/1g2hAN/zmu6)

[Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software*. https://doi.org/](http://paperpile.com/b/1g2hAN/3Ra7)[10.18637/jss.v033.i01](http://dx.doi.org/10.18637/jss.v033.i01)[.](http://paperpile.com/b/1g2hAN/3Ra7)

[Gorochowski, Thomas E., Claire S. Grierson, and Mario di Bernardo. 2018. “Organization of Feed-Forward Loop Motifs Reveals Architectural Principles in Natural and Engineered Networks.” *Science Advances* 4 (3): eaap9751.](http://paperpile.com/b/1g2hAN/AQkb)

[Greenbury, Sam F., Iain G. Johnston, Matthew A. Smith, Jonathan P. K. Doye, and Ard A. Louis. 2010. “The Effect of Scale-Free Topology on the Robustness and Evolvability of Genetic Regulatory Networks.” *Journal of Theoretical Biology* 267 (1): 48–61.](http://paperpile.com/b/1g2hAN/a0Xa)

[Guo, Yipei, and Ariel Amir. 2021a. “Exploring the Effect of Network Topology, mRNA and Protein Dynamics on Gene Regulatory Network Stability.” *Nature Communications* 12 (1): 130.](http://paperpile.com/b/1g2hAN/nEXE)

[———. 2021b. “Publisher Correction: Exploring the Effect of Network Topology, mRNA and Protein Dynamics on Gene Regulatory Network Stability.” *Nature Communications* 12 (1): 994.](http://paperpile.com/b/1g2hAN/afjs)

[Han, Heonjong, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, et al. 2018. “TRRUST v2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions.” *Nucleic Acids Research*. https://doi.org/](http://paperpile.com/b/1g2hAN/note)[10.1093/nar/gkx1013](http://dx.doi.org/10.1093/nar/gkx1013)[.](http://paperpile.com/b/1g2hAN/note)

[Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. 2010. “Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.” *PloS One* 5 (9). https://doi.org/](http://paperpile.com/b/1g2hAN/6y8v)[10.1371/journal.pone.0012776](http://dx.doi.org/10.1371/journal.pone.0012776)[.](http://paperpile.com/b/1g2hAN/6y8v)

[Larvie, Joy Edward, Mohammad Gorji Sefidmazgi, Abdollah Homaifar, Scott H. Harrison, Ali Karimoddini, and Anthony Guiseppi-Elie. 2016. “Stable Gene Regulatory Network Modeling From Steady-State Data.” *Bioengineering (Basel, Switzerland)* 3 (2). https://doi.org/](http://paperpile.com/b/1g2hAN/YjbK)[10.3390/bioengineering3020012](http://dx.doi.org/10.3390/bioengineering3020012)[.](http://paperpile.com/b/1g2hAN/YjbK)

[Lee, Tong Ihn, Nicola J. Rinaldi, François Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, et al. 2002. “Transcriptional Regulatory Networks in Saccharomyces Cerevisiae.” *Science* 298 (5594): 799–804.](http://paperpile.com/b/1g2hAN/dqjm)

[Mangan, S., and U. Alon. 2003. “Structure and Function of the Feed-Forward Loop Network Motif.” *Proceedings of the National Academy of Sciences*. https://doi.org/](http://paperpile.com/b/1g2hAN/siLg)[10.1073/pnas.2133841100](http://dx.doi.org/10.1073/pnas.2133841100)[.](http://paperpile.com/b/1g2hAN/siLg)

[Marbach, Daniel, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, et al. 2012. “Wisdom of Crowds for Robust Gene Network Inference.” *Nature Methods* 9 (8): 796–804.](http://paperpile.com/b/1g2hAN/4jdq)

[Margolin, Adam A., Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. 2006. “Reverse Engineering Cellular Networks.” *Nature Protocols* 1 (2): 662–71.](http://paperpile.com/b/1g2hAN/y84T)

[May, Robert M. 1972. “Will a Large Complex System Be Stable?” *Nature*. https://doi.org/](http://paperpile.com/b/1g2hAN/4wU4)[10.1038/238413a0](http://dx.doi.org/10.1038/238413a0)[.](http://paperpile.com/b/1g2hAN/4wU4)

[Mendes, P., W. Sha, and K. Ye. 2003. “Artificial Gene Networks for Objective Comparison of Analysis Algorithms.” *Bioinformatics*. https://doi.org/](http://paperpile.com/b/1g2hAN/nK6W)[10.1093/bioinformatics/btg1069](http://dx.doi.org/10.1093/bioinformatics/btg1069)[.](http://paperpile.com/b/1g2hAN/nK6W)

[Milo, R. 2002. “Network Motifs: Simple Building Blocks of Complex Networks.” *Science*. https://doi.org/](http://paperpile.com/b/1g2hAN/qxW1)[10.1126/science.298.5594.824](http://dx.doi.org/10.1126/science.298.5594.824)[.](http://paperpile.com/b/1g2hAN/qxW1)

[Mirzasoleiman, Baharan, and Mahdi Jalili. 2011. “Failure Tolerance of Motif Structure in Biological Networks.” *PloS One* 6 (5): e20512.](http://paperpile.com/b/1g2hAN/ybbW)

[Pratapa, Aditya, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. n.d. “Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data.” https://doi.org/](http://paperpile.com/b/1g2hAN/zcTV)[10.1101/642926](http://dx.doi.org/10.1101/642926)[.](http://paperpile.com/b/1g2hAN/zcTV)

[Prill, Robert J., Pablo A. Iglesias, and Andre Levchenko. 2005. “Dynamic Properties of Network Motifs Contribute to Biological Network Organization.” *PLoS Biology* 3 (11): e343.](http://paperpile.com/b/1g2hAN/dgp1)

[Santos-Zavaleta, Alberto, Heladia Salgado, Socorro Gama-Castro, Mishael Sánchez-Pérez, Laura Gómez-Romero, Daniela Ledezma-Tejeida, Jair Santiago García-Sotelo, et al. 2019. “RegulonDB v 10.5: Tackling Challenges to Unify Classic and High Throughput Knowledge of Gene Regulation inE. coliK-12.” *Nucleic Acids Research*. https://doi.org/](http://paperpile.com/b/1g2hAN/kB9t)[10.1093/nar/gky1077](http://dx.doi.org/10.1093/nar/gky1077)[.](http://paperpile.com/b/1g2hAN/kB9t)

[Schaffter, Thomas, Daniel Marbach, and Dario Floreano. 2011. “GeneNetWeaver: In Silico Benchmark Generation and Performance Profiling of Network Inference Methods.” *Bioinformatics*. https://doi.org/](http://paperpile.com/b/1g2hAN/PBVb)[10.1093/bioinformatics/btr373](http://dx.doi.org/10.1093/bioinformatics/btr373)[.](http://paperpile.com/b/1g2hAN/PBVb)

[Shen-Orr, Shai S., Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. “Network Motifs in the Transcriptional Regulation Network of Escherichia Coli.” *Nature Genetics* 31 (1): 64–68.](http://paperpile.com/b/1g2hAN/2eRV)

[Siahpirani, Alireza F., and Sushmita Roy. 2017. “A Prior-Based Integrative Framework for Functional Transcriptional Regulatory Network Inference.” *Nucleic Acids Research* 45 (4): 2221.](http://paperpile.com/b/1g2hAN/eLIT)

[Stone, Lewi. 2018. “The Feasibility and Stability of Large Complex Biological Networks: A Random Matrix Approach.” *Scientific Reports* 8 (1): 8246.](http://paperpile.com/b/1g2hAN/4C63)

[Tibshirani, Robert. 1996. “Regression Shrinkage and Selection Via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/](http://paperpile.com/b/1g2hAN/Kn5C)[10.1111/j.2517-6161.1996.tb02080.x](http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x)[.](http://paperpile.com/b/1g2hAN/Kn5C)

[Tjärnberg, Andreas, Torbjörn E. M. Nordling, Matthew Studham, and Erik L. L. Sonnhammer. 2013. “Optimal Sparsity Criteria for Network Inference.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 20 (5): 398–408.](http://paperpile.com/b/1g2hAN/QFDx)

[Van den Bulcke, Tim, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. 2006. “SynTReN: A Generator of Synthetic Gene Expression Data for Design and Analysis of Structure Learning Algorithms.” *BMC Bioinformatics* 7 (January): 43.](http://paperpile.com/b/1g2hAN/STjv)

[Watts, Duncan J., and Steven H. Strogatz. 1998. “Collective Dynamics of ‘small-World’ Networks.” *Nature*. https://doi.org/](http://paperpile.com/b/1g2hAN/B7wJ)[10.1038/30918](http://dx.doi.org/10.1038/30918)[.](http://paperpile.com/b/1g2hAN/B7wJ)

[———. 2011. “Collective Dynamics of ‘Small-World’ Networks.” *The Structure and Dynamics of Networks*. https://doi.org/](http://paperpile.com/b/1g2hAN/nyO2)[10.1515/9781400841356.301](http://dx.doi.org/10.1515/9781400841356.301)[.](http://paperpile.com/b/1g2hAN/nyO2)

[Young, William Chad, Adrian E. Raftery, and Ka Yee Yeung. 2014. “Fast Bayesian Inference for Gene Regulatory Networks Using ScanBMA.” *BMC Systems Biology* 8 (April): 47.](http://paperpile.com/b/1g2hAN/ZXRv)

[Zavlanos, Michael M., A. Agung Julius, Stephen P. Boyd, and George J. Pappas. 2011. “Inferring Stable Genetic Networks from Steady-State Data.” *Automatica*. https://doi.org/](http://paperpile.com/b/1g2hAN/stP0)[10.1016/j.automatica.2011.02.006](http://dx.doi.org/10.1016/j.automatica.2011.02.006)[.](http://paperpile.com/b/1g2hAN/stP0)

[Zrimec, Jan, Christoph S. Börlin, Filip Buric, Azam Sheikh Muhammad, Rhongzen Chen, Verena Siewers, Vilhelm Verendel, Jens Nielsen, Mats Töpel, and Aleksej Zelezniak. 2020. “Deep Learning Suggests That Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure.” *Nature Communications* 11 (1): 6141.](http://paperpile.com/b/1g2hAN/vzVx)

**Table 1. Transcriptional interaction graph properties**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Organism** | **# of nodes** | **% of nodes that participate in FFL-motif** | **% of FFL-motifs with shared nodesfrom the largest connected component** | **Sparsity** | **In-degree** | **Out-degree** |
| ***E. coli*** | 1917 | 37.4 | 99.1 | 2.328 | 1.106 | 1.222 |
| ***S. cerevisiae*** | 4441 | 27.0 | 100 | 2.899 | 1.421 | 1.477 |
| ***M. musculus*** | 2862 | 31.5 | 99.7 | 2.643 | 1.274 | 1.369 |
| ***H. sapiens*** | 2456 | 34.7 | 99.9 | 2.944 | 1.364 | 1.580 |

**Figure 1. Motif collection**. The five possible three-node motifs with 2 or 3 unidirectional links.

**Figure 2**. **Node participation in FFL motif.** An example of 3-node motif counts given on an FFL motif. Node *a* plays different roles in two FFL motifs ({a,b,c} and {d,a,e} respectively).

**Figure 3**. **Graphic outline of the FFLatt algorithm.** It starts with selecting a seed from the target network, and then iteratively grows the nucleus until the required size is reached. Finally, the sparsity of the network is adjusted according to the sparsity level.

**Figure 4. Pseudocode description of the FFLatt algorithm.**

**Figure 5. Attachment rules that create FFL-motif enriched network**; *p1*, *p2*, *p3*, *p4* correspond to probabilities for choosing rule *R* at next iteration while growing network. FFLTTG and FFLTTT correspond to different FFL-motif types, where *G* or *T* indicate that node has only incoming or incoming and outgoing edges respectively. The red dotted arrows here show new edges added to the network and the solid blue arrows show edges participating in the FFL motif with the new edges.

**Figure 6. Topological properties of simulated networks (*E. coli*).** FFL-motif node participation, average sparsity, in- and out-degree distribution in simulated networks. For FLL-motif node participation counts, up to three participations for each node were allowed  (in different roles).

**Figure 7. Motif enrichment analysis of 3-gene network motifs in simulated networks (*E. coli*)**. For networks generated with GNW, the *E. coli* RegulonDB [(Santos-Zavaleta et al. 2019)](https://paperpile.com/c/1g2hAN/kB9t) database was used. For networks generated with FFLatt, we used transcriptional interaction graph properties for *E. coli* specified in Table1. RandG is a random assignment of links and DAG is the same with cycles removed. NetworkX GRNs are scale-free.

**Figure 8. Stability of randomly wired simulated network graphs**. *λ* is the lowest eigenvalue of the interaction matrix M. Each data point was calculated as the average of ten different repeats of overlaying links chosen randomly with strengths from a standard distribution, with corresponding semi-transparent areas indicating the 95% confidence interval.

**Figure 9. Degree distribution in simulated networks**..