Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category
OXFORD

Genome analysis

GRN-VAE: A Simplified and Stabilized SEM Model for Gene Regulatory Network Inference

Hao Zhu [1,*] **and Donna K. Slonim** [1,*]

[1] Department of Computer Science, Tufts University, Medford, MA 02155, USA.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Many computational tools attempt to infer gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data. One recent advance is DeepSEM, a deep generative model generalizing the Linear Structural Equation Model (SEM) that improves benchmark performance over popular GRN inference methods. While DeepSEM is promising, its results are not stable over multiple runs. To overcome the instability and resolve dropout handling concerns, we propose GRN-VAE.

Results: GRN-VAE improves stability and efficiency while maintaining accuracy by delayed introduction of the sparse loss term. To minimize the negative impact of dropout in single-cell data, GRN-VAE trains on non-zero data. Most importantly, we introduce a novel idea, Dropout Augmentation, to improve model robustness by adding a small amount of simulated dropout to the data. GRN-VAE compares favorably to other methods on the BEELINE benchmark data sets, using several collections of "ground truth" regulatory relationships, and on a real-world data set, where it efficiently provides stable results consistent with literature-based findings.

Conclusions: The stability and robustness of GRN-VAE make it a practical and valuable addition to the toolkit for GRN inference from single-cell data. Dropout Augmentation may have wider applications beyond the GRN-inference problem.

Availability and implementation: Source code is available at https://bcb.cs.tufts.edu/GRN-VAE

Contact: hao.zhu@tufts.edu; donna.slonim@tufts.edu

1 Introduction

Gene regulatory networks, or GRNs, have long being used as an effective tool to represent and study the complex regulatory relationships among genes (Davidson and Levin, 2005; Karlebach and Shamir, 2008; Penfold and Wild, 2011). Understanding these interactions is crucial for gaining insight into developmental biology as well as identifying key points of regulation that may be amenable to therapeutic intervention (Emmert-Streib *et al.*, 2014).

Inferring GRNs from gene expression has long been an active research area, because gene expression captures a snapshot of the current cell state, lends insight into gene interactions, and is widely available (Mercatelli *et al.*, 2020). Methods for GRN inference can broadly be classified into three categories, respectively relying on Mutual Information, Decision Trees, or Bayesian Networks.

Mutual information measures the statistical dependence between two random variables and is used in methods such as ARACNE (Margolin *et al.*, 2006), CLR (Faith *et al.*, 2007), MRNET (Meyer *et al.*, 2007), and PIDC (Chan *et al.*, 2017). Decision Tree-based methods, such as GENIE3 (Huynh-Thu *et al.*, 2010) and GRNBoost2 (Moerman *et al.*, 2019), rely on variable importance, a metric to rank variables while creating trees. Bayesian Network-based methods model GRNs as causal inference problems. Examples include G1DBN (Lèbre (2009)) and ebdbNet (Rau *et al.* (2010)). By representing GRNs as Bayesian networks, or directed acyclic graphs (DAG), these methods account for gene interactions, approaching the problem from a network perspective.

In recent years, as single-cell RNA-sequencing (scRNA-seq) data has become more widely accessible (Svensson *et al.*, 2018), inferring GRNs from scRNA-seq data has drawn a lot of attention. ScRNA-seq allows researchers to analyze transcriptomic profiles on an individual cell level, providing a more detailed and accurate view of cellular diversity than traditional bulk RNA-seq methods. With scRNA-seq data, researchers can gain a more comprehensive understanding of gene regulation and its role in shaping cellular behavior.

However, using scRNA-seq data does introduce new challenges, which extend beyond the task of predicting GRNs. First, scRNA-seq data generally includes an excessive number of zero expression counts, often referred as "dropout." In nine scRNA-seq datasets used in Ghazanfar *et al.*, 2016, 57.7% - 92.3% of the count data are zeros. The noisy nature of scRNA-seq data makes it difficult to apply many statistical inference and machine learning methods directly. Second, compared with traditional bulk data, scRNA-seq introduces both spatial (via cell type) and temporal (via cell state or pseudo-time cell trajectory) variations (Nguyen *et al.*, 2021). How to handle these two axes is not often considered in traditional methods. Finally, as the size of scRNA-seq data sets grows, the speed and efficiency of inference algorithms becomes a limiting factor. It is challenging to apply computationally-heavy algorithms to single-cell data sets in reasonable amounts of time. One common work-around is to apply heavy filtering on genes and cells. However, filtering out too many genes or cells are filtered out limits the capacity to generate useful and novel regulatory knowledge.

Due to the high dimensional and noisy nature of single-cell data, many researchers have turned to deep neural networks as a means of addressing these challenges. Among the various network structures, variational autoencoders (VAEs) (Kingma and Welling, 2013) have garnered significant attention. A VAE consists of an encoder and a decoder, which respectively compress and reconstruct the input data through a low-dimensional latent representation. While classic autoencoders are trained using reconstruction loss alone, VAEs consider the latent variable as a random variable sampled from a parameterized distribution. As a result, VAEs are trained by optimizing the sum of the reconstruction loss and the Kullback-Leibler (KL) divergence between the posterior and prior distributions, or the evidence lower bound (ELBO). This KL divergence term in the objective function encourages a more evenly distributed latent space. This evenness helps to ensure that the values in the latent space remain meaningful.

VAEs have been widely applied in various fields, including computer vision and natural language processing, for tasks such as representation learning (encoder) and content generation (decoder). They offer several advantages, including the ability to process noisy data and no requirement for additional labeled data. These advantages make VAEs well-suited for handling noisy single-cell data. In prior work, Wang and Gu, 2018 proposed using VAEs as a dimension reduction tool and demonstrated their effectiveness in cell population clustering and visualization. Grønbech *et al.*, 2020 introduced scVAE, which uses VAEs to estimate expected gene expression values and improve cell clustering.

However, vanilla VAEs struggle to accurately infer GRNs because they do not naturally incorporate notions of causality or regulation. To support such concepts, VAEs could be transformed into a form of structural equation model (SEM), which is a class of statistical models that measures the relationships among variables. This approach was first proposed in DAG-GNN (Directed Acyclic Graph - Graph Neural Network) by Yu *et al.*, 2019. Shu *et al.*, 2021 then relaxed the acyclic graph constraint, adapting this method for GRN inference. DeepSEM reports better performance on most BEELINE benchmarks (Pratapa *et al.*, 2020) and runs significantly faster than many current methods. However, we find that its results are unstable over multiple runs, resulting in considerable performance variation and differences in the inferred networks (Fig 1b).

Here, we propose GRN-VAE which stabilizes the results of DeepSEM by only restricting the sparsity of the adjacency matrix at a later stage. It also comes with a simplified, bare-bones VAE design that removes the weight connections between encoders and decoders.

Further, we introduce the idea of dropout augmentation, or simulating a small amount of additional dropout, to improve the robustness of learning models. While this is a common approach in many other neural network applications, it appears to be novel for inferring networks from single-cell data. Here we demonstrate its utility in stabilizing inference methods.

2 Materials and Methods

2.1 Datasets Used

2.1.1 BEELINE single-cell benchmarks

In this study, we compare the performance of our proposed method, GRN-VAE, with the DeepSEM approach using the seven scRNA-seq datasets from the BEELINE benchmarks (Pratapa *et al.*, 2020). The BEELINE benchmarks consist of both synthetic expression data based on curated ground truth networks, as well as seven pre-processed real single-cell RNA-seq datasets. These scRNA-seq datasets come from both human and mouse samples and have undergone different pre-processing steps, including normalization, depending on the original data format (e.g. raw counts, transcripts per million). In some aspects, this variety reflects the wide array of differences in real-world data.

Next, BEELINE combines the scRNA-seq data with three different sources of "ground truth" data about regulatory relationships, including the functional interaction network STRING (Szklarczyk *et al.*, 2019), non-cell-type specific transcriptional networks, and cell-type specific networks. The non-cell-type specific network combines links from DoRothEA (Garcia-Alonso *et al.*, 2019), RegNetwork (Liu *et al.*, 2015), and TRRUST v2 (Han *et al.*, 2018)). The cell-type specific networks are created by the BEELINE authors for each dataset by searching through the ENCODE, ChIP-Atlas and ESCAPE databases. To generate a benchmarking dataset, BEELINE identifies highly variable transcription factors and genes and randomly samples from this pool to create a benchmark with desired size.

In our experiments, we are using the exact same evaluation dataset used in the DeepSEM paper. Note that performance of *all* methods on the non-cell-type-specific networks is little different from random, as reported in (Shu *et al.*, 2021), so we do not discuss results on those networks.

2.1.2 Hammond microglial data

To assess GRN-VAE's performance in a more practical context, we use a published data set from Hammond *et al.*, 2019 (data available from NCBI's Gene Expression Omnibus database (Edgar *et al.*, 2002) under accession GSE121654). The Hammond dataset includes RNA sequencing counts from 76,149 individual microglia in mice during multiple developmental stages / ages and after a demyelinating injury. To compare our network inference on this data set to that of DeepSEM, we compare networks from 3 adult male mice subjected to demyelinating injury caused by lysolecithin (LPC) injection to 3 adult male saline-exposed controls (samples GSM3442041 to GSM3442046).

To preprocess the data, following suggestions from Green *et al.*, 2022 for the same data, we filter out cells with below 400 or above 3000 unique genes; cells with more than 10,000 UMIs, and cells with over 3% of reads mapping to mitochondrial genes. After such filtering, there are 2,722 cells subjected to LPC treatment and 2,623 saline-treated control cells. We also apply a very light gene filter, selecting the 5,000 genes whose expression values are most variable across all six samples. We then normalize the data using natural log transformation. Note that compared to the original Hammond *et al.*, 2019 paper, here we are using a very different approach, analyzing data from all cells for a condition (LPC or control) together, instead of analyzing microglial subpopulations defined by specific injury responsive cell clusters.

2.2 Evaluation Metrics

Following the suggestions from the BEELINE paper, we choose Early Precision Ratio (EPR) as our primary evaluation metric. Early Precision is "the fraction of true positives in the top-k edges," where k is number of edges in the "ground truth" network. EPR compares the ratio of the measured Early Precision against the performance of a random predictor.

2.3 Model structure

2.3.1 SEM style VAE

Let $X \in \mathbb{R}^{|c| \times |g|}$ be a gene expression matrix from a single-cell experiment, where $|c|$ is the number of cells and $|g|$ is the number of genes. The task of GRN inference is to infer a weighted adjacency matrix $A \in \mathbb{R}^{|g| \times |g|}$ based on the expression data X. Similar to many Bayesian Network methods, the SEM style VAE (Yu *et al.*, 2019; Shu *et al.*, 2021) starts by making a linear additive assumption that can be written as

$$X = XA + Z, \tag{1}$$

where $Z \in \mathbb{R}^{|c| \times |g|}$ is a random variable describing the variance. In other words, this equation simply states that we assume the amount of expression of any genes equals to the weighted sum of all genes that regularize it. Starting from here, by rearranging the terms, we can easily get the following two equations

$$Z = X(I - A), \tag{2}$$

$$X = Z(I - A)^{-1}. \tag{3}$$

Equation 2 infers the random variable Z from X and Equation 3 is a generative model that reconstructs X based on random variable Z. These two equations naturally fit into a VAE framework with Equation 2 as an encoder and Equation 3 as a decoder. In this case, Z is the latent random variable. For a VAE, the problem of finding the set of parameters θ that maximizes the log evidence $\log(P(X))$ is intractable. Instead, people often maximize the evidence lower bound (ELBO), which we write as

$$ELBO = -D_{KL}(q(Z|X)||p(Z)) + E_{z \sim q(Z|X)}[\log p(X|Z)], \tag{4}$$

where the first term is the KL divergence and the second term is the reconstruction loss.

The authors of DeepSEM also introduced an L1 sparse loss term that regulates the sparsity of the learned adjacency matrix, and two hyperparameters, α and β, to control the influence of the sparse loss and KL divergence. The final form of the objective function is to minimize the following loss function:

$$\begin{aligned} Loss = & -E_{z \sim q(Z|X)}[\log p(X|Z)] \\ & +\beta D_{KL}(q(Z|X)||p(Z)) \\ & +\alpha||A||. \end{aligned} \tag{5}$$

The SEM style VAE can be viewed as a neural network with two separate components: a parameterized adjacency matrix, and a neural network designed for learning and reconstructing the features of genes. During the training stage, these two networks are trained in an alternating order with two separate optimizers.

After a model converges, we extract the adjacency matrix from the model and convert it to an adjacency list. This list is then sorted based on the absolute value of the edge weights. Note that positive edge weights correspond to up-regulation or stimulation, while negative edge weights correspond to down-regulation or inhibition.

2.3.2 Structural Differences between GRN-VAE and DeepSEM

Using the same framework and loss function, GRN-VAE and DeepSEM are similar in terms of model design, but they have the following difference. DeepSEM estimates a separate latent variable Y representing the prior of the latent variable, while GRN-VAE simply assumes that the prior is normally distributed. The benefit of this difference is that with the same set of hyperparameters, we reduce the number of parameters in the model by 73.2%. There is also a closed-form solution for the KL divergence. Compared to DeepSEM, GRN-VAE is structured more like a classic VAE model and is much easier to implement.

3 Experiments and results

In this section, we start by improving model stability when the model is trained on the entire expression data set, as recommended in the DeepSEM paper. Then, due to concerns about dropout, we discuss why and how to train the model on just the non-zero data while keeping our stability gains. Finally, using the Hammond data, we demonstrate the importance of our work on improving model stability in generating scientifically meaningful insights.

3.1 Improved stability when trained on entire dataset

3.1.1 Instability in Original DeepSEM

Although DeepSEM yields impressive performance compared to other GRN inference methods (Shu *et al.*, 2021), we found that its results were not stable across replicated runs. As an example, on the hESC datasets with TFs + 500 genes validated on the STRING network, the average EPR across 10 runs reported in the BEELINE paper was 4.13. Here, we repeat the DeepSEM algorithm 100 times on the same dataset. We find that the EPR in 14 out of the 100 runs is smaller than 3.70, as shown in Figure 1a. The authors of DeepSEM acknowledged this variation, but suggested that using a different optimizer would solve this issue. However, we found that the instability persisted regardless of the choice of optimizer (data not shown). Further investigation showed that cases with low EPR performance all also have low sparse loss at the point of convergence, as shown in Figure 1b. We therefore hypothesized that when the minimization of the sparse loss is prioritized, the predicted adjacency matrix would be less precise.

3.1.2 Delayed Addition of the Sparse Loss Stabilizes Performance

As shown in Equation 5, the L1 sparse loss helps to regularize parameters on the adjacency matrix and prevents it from getting too noisy. However, it may also lead the model to converge to a local minimum during optimization. Here we suggest a simple approach that delays the addition of the sparse loss term for a number of iterations. In this way, the model will have some time to "warm up." Broadly, this technique falls into the category of progressive curriculum learning (Soviany *et al.*, 2022; Wang *et al.*, 2021), where certain parts of the models are progressively updated to guide the model from an easier task to a more complicated task. In Figure 2, we demonstrate that with as few as 10 delayed iterations before adding the sparse loss, we see reduced variance in the results on all 7 BEELINE benchmarks using TFs + 500 genes evaluated on the STRING network (with similar trends observed for the other BEELINE benchmarks). The ideal number of delayed steps may depend on the number of samples/cells in the data. For a typical scRNA-seq data set, overall we recommend delaying the addition of the sparse loss by ∼10-30 iterations.

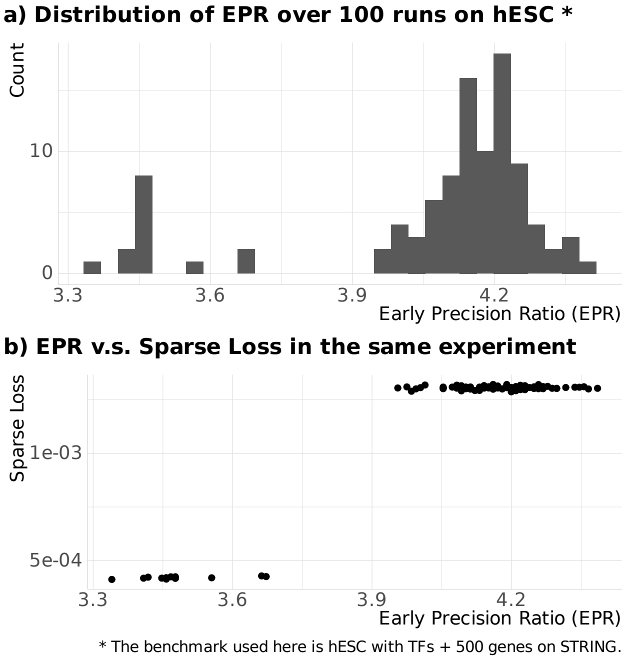


Fig. 1. a. Original DeepSEM is not stable, producing a wide range of EPR values in repeated runs on the same data set. b. Poorer performance appears to be correlated with the sparse loss.

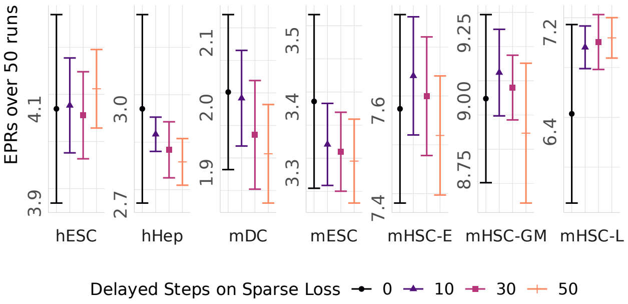


Fig. 2. Delayed sparse losses reduce performance (EPR) variances in 50 runs on the BEELINE benchmarks (TF + 500 genes) based on the STRING network

3.2 Improved dropout handling

3.2.1 Training Model on Non-Zero Data Only

Single-cell data is well-known to be noisy. Importantly, many counts of zero in scRNA-seq data do not truly reflect that there is no expression of the corresponding transcripts. Rather, these are readings missed by the instruments, especially when the true expression counts are low. Such zeros are referred to as "dropout."

The VAE network structure can handle a certain level of noise, but when the noise is so prevalent and systematic, model performance will be impacted because we are forcing the model decoder to generate more random zeros than it truly should. One way to handle dropout is to train the autoencoder on only the non-zero counts. In other words, when we account for loss, we ignore prediction errors on fields where the original expression counts are reported as zeros. In this way, all the numbers the model encounters are real, which should theoretically improve model performance.

However, in practice, we find that this method often hurts model performance on many benchmarks and introduces considerable variance in the results, as shown in Figure 3. We suspect a side effect of focusing on non-zero values is that it makes the model less robust and more sensitive to random noise or other factors such as sparse loss. Which effect, the beneficial or harmful one, is observed on a data set may depend on the specific number and distribution of zero counts in each dataset. We note that hESC, hHep, and mHSC-L happen to be the three datasets with the highest percentage of zeros. Dropping the zero-count data in such cases may lead to substantial bias in the model. Another possible reason for the differences might be the different normalization methods used to preprocess the data (Pratapa *et al.*, 2020).

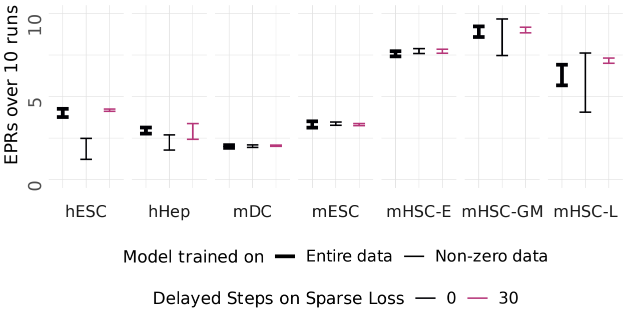


Fig. 3. Training the model on non-zero data may make the model less robust. Delayed sparse loss helps in many cases, but not always. Results reported over 10 runs on the BEELINE benchmarks (TF + 500 genes) based on the STRING network.

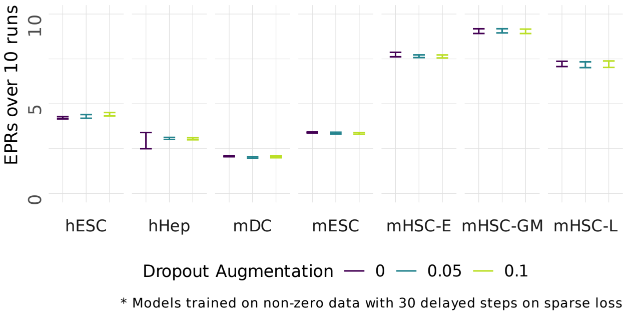


Fig. 4. Dropout augmentation and delayed addition of sparse loss make it possible to train with non-zero data only. Results reported in 10 runs on the BEELINE benchmarks (TF + 500 genes) based on the STRING network.

As shown in Figure 3, we discovered that introducing a delayed sparse loss often helps to resolve the negative impact of training only on non-zero counts. However, in other cases, it either has no effect or still leads to large performance variance (e.g., sample hHep in Figure 3). This observation encouraged us to think about other methods to improve model robustness.

3.2.2 Dropout Augmentation Keeps Models Robust

We hypothesized that further adding a small number of random zeros, as if simulating additional dropout, could improve model robustness. We call this method "dropout augmentation" (DA). This idea follows from a simple assumption: if some of the original dropout in scRNA-seq data occurs at random, then a robust model should provide similar results even if we augment the data with some additional dropout. A random dropout distribution is both widely assumed and occasionally contradicted in the scRNA-seq literature. Here, we don’t actually require it to be fully correct, but it serves to inspire the DA approach.

Our findings in Figure 4 show that an additional $5-10\%$ of DA can help stabilize the model and enable successful training on non-zero data only. We also notice that in most cases, a small percentage of DA improves model performance. However, beyond some point, as we increase the amount of dropout, convergence time may increase and performance degrade because

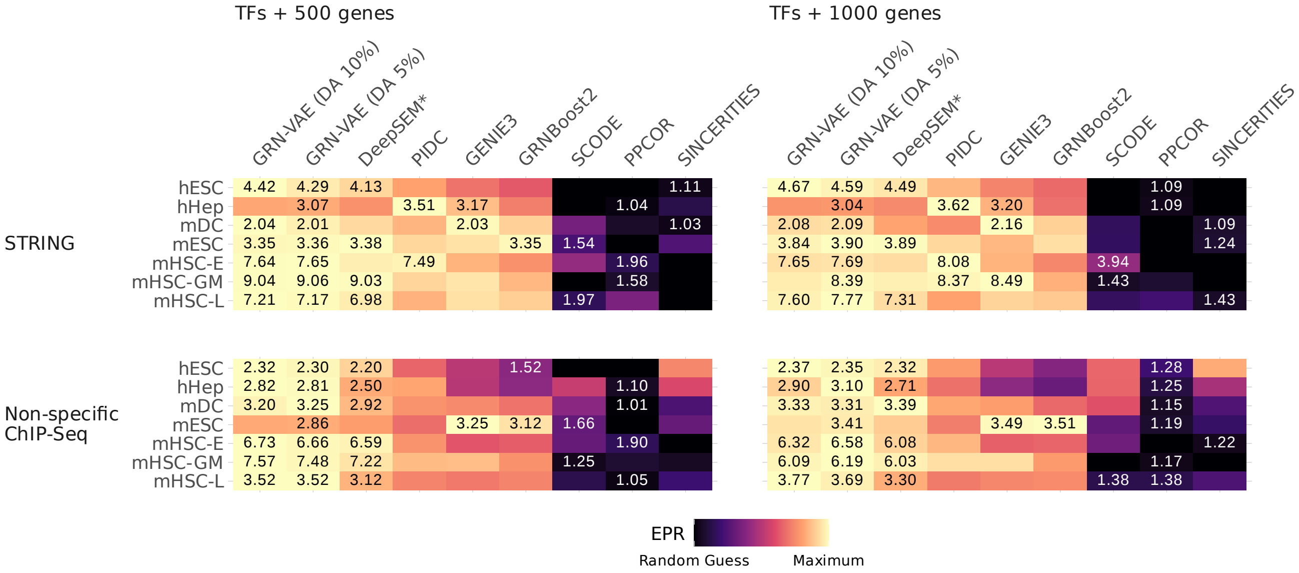


Fig. 5. Comparison of GRN-VAE's performance vs several comparator methods over BEELINE benchmarks. Comparator performance come from the DeepSEM paper.

too much noise makes it too difficult to learn. We have not yet identified a consistent pattern across datasets, but we found that 5-10% DA works in most cases.

3.2.3 Comparison of GRN-VAE to Prior Methods

We choose this final model with delayed loss, training on non-zero data, and dropout augmentation, to compare to prior methods in GRN inference on the BEELINE benchmarks. A complete comparison between GRN-VAE and the other algorithms assessed in (Shu *et al.*, 2021) is shown in Figure 5. Note that for this comparison, we are using the same data as reported in the Shu *et al.* (2021) paper, and the results for all the comparitor methods except for the two versions of GRN-VAE are taken from the supplement of the DeepSEM paper.

GRN-VAE does not always have the highest EPR, but it is frequently highest and always close to the best, where most other methods are either consistently worse or have much more variable performance depending on the data set. In particular, GRN-VAE's EPR is quite close to DeepSEM's.

3.3 Variational Visualization of the latent space in GRN-VAE

The KL divergence helps a VAE enforce having a smooth and meaningful latent space, such that when we randomly sample from the learned distribution Z, we can always generate a truthful reconstruction. In Figure 6, we compare the UMAP (McInnes *et al.*, 2018) visualization of the original expression data of hESC in BEELINE with the learned latent variable from GRN-VAE on the same data.

Although UMAP successfully separates cells into temporal clusters in both cases, the later one has the advantage of showing the trajectory as a clear gradient. From the perspective of representation learning, the large blank space in the first plot does not have a practical meaning, suggesting that the relative positions of the clusters are difficult to interpret. In contrast, the points in the second plot are evenly distributed while the clusters remain clearly separated. Thus, the trajectory line observed in the GRN-VAE case is much smoother, although it still features the characteristic "arch" shown by many such projection methods on data including a continuous latent variable such as time (Diaconis *et al.*, 2008). At the same time, since we learned a probabilistic distribution on the latent variable, the variations and uncertainties of the learned model could also be visualized by repeatedly sampling from the learned distribution.

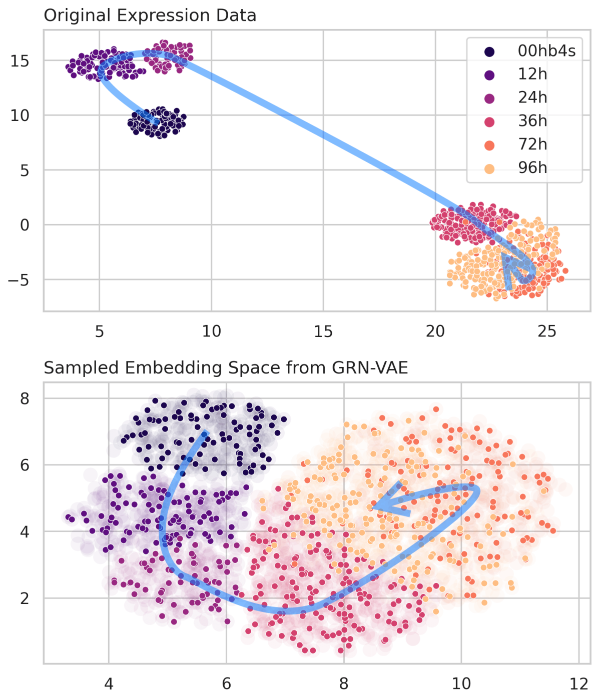
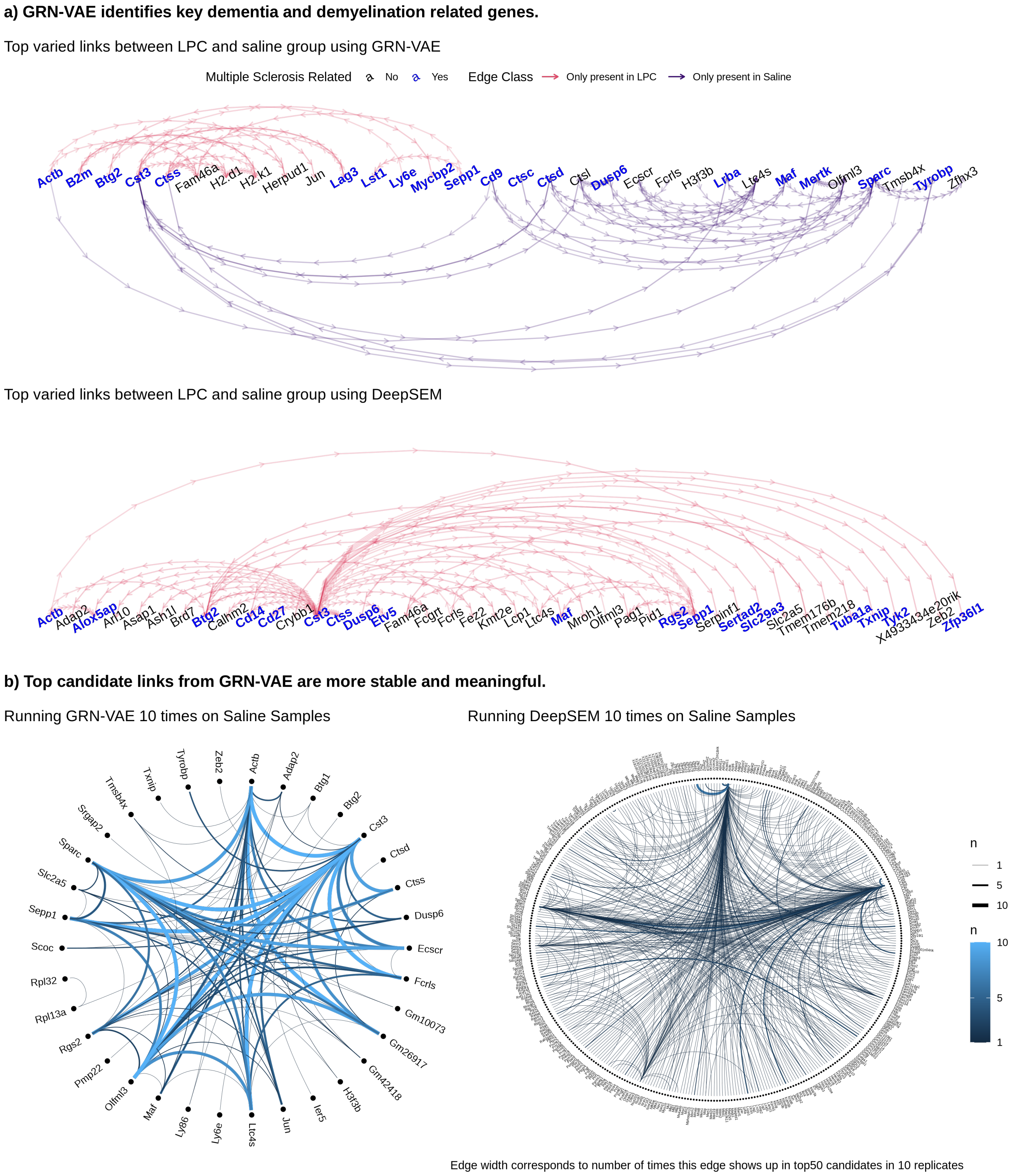


Fig. 6. UMAP visualizations of the original hESC expression data and the learned latent variable in GRN-VAE. The uncertainties in the lower plot are collected by sampling the latent distribution three times.

3.4 GRN-VAE characterizes microglial dysregulation

To evaluate how GRN-VAE could help generate scientific insights in a real-world research scenario, we assess its ability to identify affected regulations under induced perturbations. We apply both GRN-VAE and DeepSEM independently on cells with LPC-injected demyelinating injury


a) GRN-VAE identifies key dementia and demyelination related genes.

Top varied links between LPC and saline group using GRN-VAE

Top varied links between LPC and saline group using DeepSEM

b) Top candidate links from GRN-VAE are more stable and meaningful.

Running GRN-VAE 10 times on Saline Samples

Running DeepSEM 10 times on Saline Samples

Edge width corresponds to number of times this edge shows up in top50 candidates in 10 replicates

Fig. 7. a. Top 50 most changed regulations based on GRN-VAE and DeepSEM. Inferred links on ribosomal proteins, gene models, and mitochondrial proteins were removed prior to ranking; b. Top 50 inferred regulations in 10 replicates based on predicted edge weight. An enlarged version of the DeepSEM + Saline graph is included in the online supplement.

and saline-injected control microglia from the Hammond dataset. As stated in the Methods section, here we use a fixed set of the 5,000 most-varying genes identified in data from all male mice to ensure the differences we find are meaningful.

A comparison of the top network differences between LPC-induced demyelinating injury and controls for both GRN-VAE and DeepSEM is displayed in Figure 7a. This figure simply shows the inferred regulatory relationships that appear to change most between the conditions, for each method. The figure highlights genes that have previously been linked to to Multiple Sclerosis, a demyelinating condition often modeled by LPC injection. Overall, 59% (19 out of 32) of the connected genes predicted by GRN-VAE have been previously reported as biomarkers for Multiple Sclerosis, while only 43% (18 out of 42) of the top genes predicted by DeepSEM are similarly implicated in the literature. We should also note that several "unrelated" genes in this list have been identified as biomarkers for dementia, for which demyelination itself is suggested to be a biomarker (Bouhrara *et al.*, 2018). These results suggest that, while many potential underlying relationships may not yet have been validated, a higher fraction of GRN-VAE's inferred changes are consistent with current knowledge.

To further demonstrate the impact of methodological instability on analytical results, we then repeated each of these four network inference experiments (GRN-VAE on LPC-treated animals, GRN-VAE on Saline-treated animals, DeepSEM + LPC, and DeepSEM + Saline) ten times. The results of this stability test are visualized in Figure 7b. Here, the top 50 predicted links are extracted from each replicate based on the absolute value of the adjacency matrix; the direction of the change in regulation is not indicated. We simply count how often the links found in the top 50 results appear across the ten replicates. We see that most edges are consistently repeated across the GRN-VAE replicates. However, for DeepSEM, the most frequent edge occurred in just half of the replicates, and the vast majority of edges were seen only once. This instability illustrates that crucial regulatory relationships may be missed by an unstable inference method.

3.5 Runtime Analysis

The GRN-VAE algorithm runs on the entire adjacency matrix, resulting in time complexity quadratic in the number of genes and linear in the number of cells. More pragmatically, the GRN-VAE algorithm on one BEELINE benchmark with 910 genes and 758 cells converges in 15.9 seconds on a Nvidia A100 GPU. This is a significant improvement compared to the 43.2 seconds' execution time using DeepSEM, thanks to our optimized implementation and the 70% reduction in parameters described in the methods section.

Furthermore, when applied to the large Hammond dataset, which includes 5,000 genes and over 2,600 cells, our algorithm finished within 10 minutes after 500 iterations. GRN inference on such large datasets is simply not computationally feasible for many prior algorithms without heavy filtering or clustering. This analysis thus demonstrates the scalability of GRN-VAE and its capacity to be applied on large real-world datasets.

4 Discussion

4.1 Rethinking dropout modeling in scRNA-seq data

Dropout is an interesting pattern inherent in scRNA-seq data. On the one hand, it makes modeling single-cell data a challenging task. On the other hand, evidence suggests that dropout is more likely on low counts, suggesting that there is some information in it. It is even possible to identify functional information by clustering cells' dropout patterns (Qiu, 2020).

Nevertheless, instead of eliminating dropout through any form of imputation, the dropout augmentation idea we propose in this paper attempts to solve this problem by paradoxically adding more noise to the data. In this way, our model becomes more robust and is essentially "vaccinated" against dropout. Dropout augmentation is a simple concept that can be easily adopted for many other single-cell methods. We are not aware of any prior work using this approach on single-cell data, but similar strategies are used in many other fields, such as computer vision and voice recognition (Li *et al.* (2019); Liu *et al.* (2020)). In addition to adding noise to input data, another strategy to introduce dropout robustness is to randomly set some model parameters to zero. This is a very popular technique that prevents model overfitting in deep learning (Srivastava *et al.*, 2014).

4.2 Stability & robustness are crucial for scientific insights

The stability and robustness of an algorithm may not seem exciting, but they are crucial for generating meaningful scientific insights. Without consistent and reliable results, a model not only undermines the validity of scientific findings but also makes it difficult to generate new knowledge through comparison. In this work, we focus on this aspect of GRN inference methods. Our results on the Hammond dataset clearly demonstrate the importance of model stability.

4.3 Limitations

Through the BEELINE benchmarks, we discovered that dropout augmentation can enhance a model's performance up to a certain point. Beyond that threshold, added noise can impact the model's ability to learn from data and ultimately harm its performance. We also find that different data sets have varying thresholds. For example, a 30% augmentation ratio improves the average EPR performance for hESC + STRING with TFs + 500 genes to 4.78, a 16% increase over DeepSEM. However, a 30% augmented dropout harms the method's performance on the mHSC-E and mHSC-GM data sets, which seem to do better with a lower augmentation percentage around 5%. Future work should investigate ways to predict a suitable value for a given data set. We have observed a correlation between the percentage of zeros and the optimal dropout ratio, but more work would be needed to demonstrate a causal relationship. Therefore we suggest keeping the augmentation percentage as a hyperparameter and recommend a default value of 5-10%.

5 Conclusion

GRN-VAE simplifies the most successful and fastest of recent GRN inference methods for single-cell RNA-sequencing data. While maintaining comparable accuracy on benchmark data sets, our method demonstrates improved robustness, stability, and efficiency. Further, the incorporation of Dropout Augmentation (DA) for neural network algorithms learning from single-cell data may have wider applications beyond GRN inference. Characterizing how best to choose the appropriate level of DA will be essential to enabling such approaches.

Acknowledgements

We thank Hantao Shu, Jianyang Zeng, and Jianzhu Ma for sharing their data and code from the original DeepSEM paper.

References

Bouhrara, M. *et al.* (2018). Evidence of demyelination in mild cognitive impairment and dementia using a direct and specific magnetic resonance imaging measure of myelin content. *Alzheimer's & Dementia*, **14**(8), 998–1004.

Chan, T. E. *et al.* (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, **5**(3), 251–267.

Davidson, E. and Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of Sciences*, **102**(14), 4935–4935.

Diaconis, P. *et al.* (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, **2**(3), 777–807.

Edgar, R. *et al.* (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**(1), 207–10.

Emmert-Streib, F. *et al.* (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, **2**, 38.

Faith, J. J. *et al.* (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, **5**(1), e8.

Garcia-Alonso, L. *et al.* (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, **29**(8), 1363–1375.

Ghazanfar, S. *et al.* (2016). Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. *BMC systems biology*, **10**(5), 11–24.

Green, L. A. *et al.* (2022). The embryonic zebrafish brain is seeded by a lymphatic-dependent population of mrc1+ microglia precursors. *Nature Neuroscience*, **25**(7), 849–864.

Grønbech, C. H. *et al.* (2020). scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, **36**(16), 4415–4422.

Hammond, T. R. *et al.* (2019). Single-cell rna sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. *Immunity*, **50**(1), 253–271.

Han, H. *et al.* (2018). Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, **46**(D1), D380–D386.

Huynh-Thu, V. A. *et al.* (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, **5**(9), e12776.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, **9**(10), 770–780.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lèbre, S. (2009). Inferring dynamic genetic networks with low order independencies. *Statistical applications in genetics and molecular biology*, **8**(1).

Li, B. *et al.* (2019). Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, **32**.

Liu, X. *et al.* (2020). How does noise help robustness? explanation and exploration under the neural sde framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 282–290.

Liu, Z.-P. *et al.* (2015). Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**.

Margolin, A. A. *et al.* (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7-1, pages 1–15. BioMed Central.

McInnes, L. *et al.* (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mercatelli, D. *et al.* (2020). Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1863**(6), 194430.

Meyer, P. E. *et al.* (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, **2007**, 1–9.

Moerman, T. *et al.* (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, **35**(12), 2159–2161.

Nguyen, H. *et al.* (2021). A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, **22**(3), bbaa190.

Penfold, C. A. and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface focus*, **1**(6), 857–870.

Pratapa, A. *et al.* (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, **17**(2), 147–154.

Qiu, P. (2020). Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, **11**(1), 1–9.

Rau, A. *et al.* (2010). An empirical bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology*, **9**(1).

Shu, H. *et al.* (2021). Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, **1**(7), 491–501.

Soviany, P. *et al.* (2022). Curriculum learning: A survey. *International Journal of Computer Vision*, pages 1–40.

Srivastava, N. *et al.* (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15**(1), 1929–1958.

Svensson, V. *et al.* (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, **13**(4), 599–604.

Szklarczyk, D. *et al.* (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**(D1), D607–D613.

Wang, D. and Gu, J. (2018). Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, **16**(5), 320–331.

Wang, X. *et al.* (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, Y. *et al.* (2019). Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR.