Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category
OUP_First_SBk_Bot_8401-eps-c

Genome analysis

Improving Gene Regulatory Network Inference using Dropout Augmentation

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 methods have been proposed to infer gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data. One major challenge when working with single-cell data is the prevalence of dropout events, when some expression values are not captured. Here we propose to improve GRN inference using Dropout Augmentation (DA), which helps models stay robust against stochastic dropout events by introducing more dropout noise during training.
Results: Benchmarking experiments illustrate the improved performance and increased stability of the proposed DAZZLE model over existing approaches. Exploration of the source of gain shows that DA plays a significant role in improving performance, while restricting the sparsity of the adjacency matrix at a later stage stabilizes the model. The practical application of the DAZZLE model on a longitudinal mouse microglia dataset, containing over 15,000 genes, illustrates its ability to handle real-world single cell data with minimal gene filtration.
Conclusions: The improved robustness and stability of DAZZLE 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: Project website: https://bcb.cs.tufts.edu/DAZZLE; Visualization of inferred mouse microglia network: https://bcb.cs.tufts.edu/DAZZLE/hammond.html; Code available at https://github.com/TuftsBCB/dazzle and on PyPI under the grn-dazzle package.
Contact: hao.zhu@tufts.edu; donna.slonim@tufts.edu

1 Introduction

Gene Regulatory Network (GRN) inference from expression data offers a contextual model of the interactions between genes *in vivo*. (Davidson and Levin, 2005; Karlebach and Shamir, 2008; Penfold and Wild, 2011). Understanding these interactions is crucial for gaining insight into development, pathology, and key points of regulation that may be amenable to therapeutic intervention (Emmert-Streib *et al.*, 2014).

While GRN inference from bulk transcriptomic data has a long history, many recent studies consider the contextual specificity offered by single-cell RNA sequence data (scRNA-seq) (Svensson *et al.*, 2018). Single cell RNA sequencing allows researchers to analyze transcriptomic profiles of individual cells, providing a more detailed and accurate view of cellular diversity than traditional bulk methods. Among established network inference methods, GENIE3 (Huynh-Thu *et al.*, 2010) and GRNBoost2 (Moerman *et al.*, 2019) are tree-based approaches initially proposed for bulk data that have been found to work well on single-cell data without modification. LEAP (Specht and Li, 2017) estimates pseudotime to infer gene co-expression over several lagged windows. SCODE (Matsumoto *et al.*, 2017) applies a similar pseudotime idea, combined with ordinary differential equation (ODE) to model the results. SCENIC (Aibar *et al.*, 2017) starts by identifying gene co-expression modules using GENIE3/GRNBoost2, followed by identifying key transcription factors (TFs) that regulate these modules or regulons. scMTNI (Zhang *et al.*, 2023) studies GRNs in different cell clusters using a multi-task learning framework. NetREX-CF (Wang *et al.*, 2022) makes optimizations based on prior GRN networks and uses collaborative filtering to address the

incompleteness of prior data. PANDA (Schlauch *et al.*, 2017) further optimizes prior GRN networks using massage passing.

The application of neural networks (NNs) in the analysis of single-cell data has advanced rapidly in the last couple of years. A leading NN-based GRN inference method, DeepSEM, (Shu *et al.*, 2021) parameterizes the adjacency matrix and uses a variational autoencoder (VAE) (Kingma and Welling, 2013) architecture optimized on reconstruction error. In fact, when compared on the BEELINE benchmarks where the "right" networks are (approximately) known, (Pratapa *et al.*, 2020), DeepSEM reports better performance than other methods and runs significantly faster than most.

Beyond GRN inference, DCA (Eraslan *et al.*, 2019), SAUCIE (Amodio *et al.*, 2019), and scGen (Lotfollahi *et al.*, 2019) all use autoencoder or VAE designs, but have different focuses. DCA prioritizes removing noise, amplification, and dropout events from single cell data. SAUCIE trains its network with multiple objectives and scGen focuses on predicting perturbation. Some major challenges in GRN inference from single-cell data (Pratapa *et al.*, 2020) include handling cellular diversity, accounting for inter-cellular variation in sequencing depth, cell-cycle problems, and dropout events. The latter is the main focus of this paper.

Dropout events describe the situation when transcripts, often those with low or moderate expression in a cell, are not counted by the sequencing technology, resulting in single-cell data sets having an excessive number of zero expression counts. This phenomenon of excessive numbers of zeros, or "zero-inflation", is known to be a primary characteristic of single-cell data. For example, in nine representative datasets examined in Ghazanfar *et al.*, 2016, 57.7% - 92.3% of the count data are zeros. Typical solutions to handle dropout events include attempting to impute the missing data, focusing only on the most variable genes, or modeling the data to account for dropout and cellular heterogeneity. With dropout imputation, there are concerns about imputation accuracy and inflated variability, although some imputation methods claim to suffer less (Chen and Zhou, 2018). Focusing only on genes with the most variable expression (or otherwise filtering the gene set) can omit crucial nodes from the regulatory network.

Other common solutions from the machine learning field have, to our knowledge, rarely been explored in the field of single-cell GRN inference. One possible route is to improve model robustness through noise injection. It has long been known that by adding noise to the input data during model training, we can improve the robustness, and sometimes even the performance, of many machine learning models. Bishop first pointed out that adding noise is equivalent to Tikhonov regularization (Bishop, 1995). Hinton further introduced the idea of using random "dropout" on either input or model parameters to improve training performance (Hinton *et al.*, 2012). In this case, adding zeros regularizes the models, improving robustness. Moreover, given that our single cell data sets are already filled with zeros, arguably due to stochastic events (though some credit biological variability), it seems reasonable to assume that a functioning model would still work with a few extra zeros.

Based on these considerations, here we propose to use "dropout augmentation" (DA), in which we add random zeros to the input single cell data during training to improve the performance of GRN inference. In this work, we apply DA to a model inspired by DeepSEM (Shu *et al.*, 2021), which, as explained above, parameterizes the adjacency matrix and models GRNs with a VAE framework. We also improve the stability of the model by only restricting the sparsity of the adjacency matrix at a later stage, and we explore the feasibility of training only on the non-zero data. In addition, we simply the model by reducing the number of parameters up to 73% by using a closed-form normally distributed prior instead of a separated latent variable. Together, we name our method DAZZLE (Dropout Augmentation for Zero-inflation Learning Enhancement). We demonstrate DAZZLE's performance both on the BEELINE benchmarks and on a real, time-series mouse microglia data set. In this latter context, we illustrate how DAZZLE's network inference facilitates interpreting typical-sized data sets efficiently, in this case explaining microglial dynamics across the mouse lifespan.

2 Materials and Methods

2.1 Datasets Used

2.1.1 BEELINE single-cell benchmarks

In the main text, we compare the performance of DAZZLE, our proposed method, primarily to the DeepSEM approach, using the seven scRNA-seq datasets from the BEELINE benchmarks. DeepSEM has previously been shown to outperform GENIE3 (Huynh-Thu *et al.*, 2010), GRNBoost2 (Moerman *et al.*, 2019), PIDC (Chan *et al.*, 2017), SCODE (Matsumoto *et al.*, 2017), ppcor (Kim, 2015), and SINCERITIES (Papili Gao *et al.*, 2018) on the same data - these comparisons are reported in the Supplement.

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 (Pratapa *et al.*, 2020). 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 or processed data). In some aspects, this variety reflects the wide array of differences we encounter 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 represented in the STRING database (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 were created by the BEELINE authors for each dataset by searching through the ENCODE (ENCODE Project Consortium, 2012), ChIP-Atlas (Zou *et al.*, 2022), and ESCAPE (Xu *et al.*, 2013) databases. To generate a benchmarking dataset, BEELINE identifies highly variable transcription factors and genes and randomly samples from this pool to create a benchmark of the desired size.

In our experiments, we use the exact evaluation dataset from the DeepSEM paper. We observed that the performance of *all* methods on the non-cell-type-specific networks was little different from random, as reported in (Shu *et al.*, 2021), so we do not further discuss results for those networks.

2.1.2 Hammond microglial data

To assess DAZZLE'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 number GSE121654). The Hammond mouse microglial dataset includes RNA sequencing counts for cells underlying several possible comparisons. In our analysis, we selected the data from five mouse developmental stages, namely embryonic day (E14.5), early postnatal day (P4/5), late juvenile stage (P30), adulthood (P100), and old age (P540). Each stage includes single cell data from four male mice.

To preprocess the data, following suggestions from Green *et al.*, 2022 for the same data, we filtered out cells with fewer than 400 or more than 3000 unique genes, cells with more than 10,000 UMIs, and cells with over 3% of reads mapping to mitochondrial genes. Many standard analysis approaches further reduce the data set size by filtering the gene set, only including the most variable genes. However, here, we only remove genes with a raw count of zero transcripts detected in all cells. We further removed all genes with only "model" RefSeq entries (genes predicted only by automated annotation pipelines, with "XM_" RefSeq prefixes (O'Leary *et al.*, 2016)), mitochondrial genes, and ribosomal genes from this pool, to

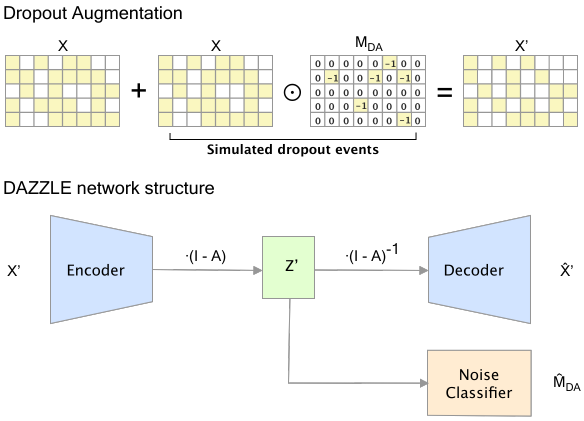

Fig. 1. Dropout augmentation and network structure

simplify the interpretation of the resulting networks. The expression values are normalized using natural log transformation. Note that compared to the original Hammond *et al.*, 2019 paper, here we are using a very different approach, analyzing changes in potential regulatory links across time, rather than attempting to identify 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. The EPR compares the ratio of the measured Early Precision against the performance of a random predictor. In addition, we also calculate the area under the precision recall curve (AUPR) and the corresponding ratio (AUPRR) as secondary measures.

2.3 Dropout Augmentation

Previous research suggests that the zero values in single cell data include both real biological zeros, corresponding to truly absent genes, and random dropout events. A successful single-cell model should remain robust regardless of how dropout values are distributed. This idea informs the dropout augmentation algorithm. 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. We randomly sample a proportion of data and temporarily replace these values with zeros. Alternatively, the augmented data could be treated as the sum of the original expression data X and a dropout noise term E, where E is the Hadamard product of negative X and a masking term derived from binomial sampling. By denoting the mask of dropout augmentation as M_{DA} and the probability of augmented dropout by p, we can write the dropout noise term E and the augmented data X' as follows:

$$E = -X \odot M_{\text{DA}}, \text{where } M_{\text{DA}}^{ij} \sim \text{Bernoulli}(p) \tag{1}$$

$$X' = X + E. \tag{2}$$

During the entire training process, M_{DA} is re-sampled every iteration, so the augmented data X' changes in every iteration. The model only sees X', not X.

2.4 GRN Inference with DAZZLE

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. In previous work, DAG-GNN (Yu *et al.*, 2019) and DeepSEM (Shu *et al.*, 2021) start by making a linear additive assumption that can be written as

$$X = XA + Z, \tag{3}$$

where $Z \in \mathbb{R}^{|c|\times|g|}$ is a random variable characterizing the noise, essentially describing the gap between the overall expected counts of the genes based on their regularizers (XA) and the observed counts (X). Here, to embrace the idea that the observed data are noisy due to dropout, we modify this assumption and rewrite equation 3 as following. Since dropout is so prevalent in single-cell data, we believe equation 4 describes the reality in a more accurate manner:

$$X' = X'A + Z', \tag{4}$$

where Z' is defined for X' analogously to the definition of Z for X. Following a similar transformation to that done in DAG-GNN and DeepSEM, by rearranging the terms, we can rewrite equation 4 in the following two forms

$$Z' = X'(I - A), \tag{5}$$

$$X' = Z'(I - A)^{-1}. \tag{6}$$

Equation 5 infers Z' from X' and Equation 6 is a generative model that reconstructs X' based on the noise sum. These two equations naturally fit into a VAE framework with Equation 5 as an encoder and Equation 6 as a decoder. When we parameterize both the VAE model and the adjacency matrix, the encoder could be denoted as $q_\phi(Z'|X')$ and the decoder could be denoted as $p_\theta(X'|Z')$, with A being part of ϕ and θ. In this case, Z' is the latent 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

$$\begin{aligned} ELBO = &- D_{\text{KL}}(q_\phi(Z'|X')||p_\theta(Z')) \\ &+ \mathbb{E}_{z'\sim q_\phi(Z'|X')}[\log p_\theta(X'|Z')], \end{aligned} \tag{7}$$

where the first term is the KL divergence and the second term can be thought as the reconstruction loss.

The random variable Z' describes the deviation of the observed value X' from the expected value $X'A$. In a particular cell, if the expression value of a particular gene happens to be observed as 0 due to dropout events, we will more likely see a larger deviation Z'. In other words, Z' contains information that could be used to infer whether a value comes from dropout events. Following this rationale, we can add a classifier C_{DA} based on the specified dropout augmentation masked M_{DA}, as shown in equation 8 below. As a naïve approach, here we choose a simple 3-layer multi-layer perceptron (MLP) to do the classification.

$$\hat{M}_{\text{DA}} = \text{sigmoid}(C_{\text{DA}}(Z')). \tag{8}$$

The loss function of this classifier is simply a binary cross entropy function.

$$L_{BCE} = -\mathbb{E}\left[M_{\text{DA}} \log \hat{M}_{\text{DA}} + (I - M_{\text{DA}}) \log(I - \hat{M}_{\text{DA}})\right] \tag{9}$$

The dropout augmentation classifier could be trained either separately or together with the main model using the same optimizer. In our experiment, we add the classification loss to the ELBO function scaled by a hyper parameter χ. Furthermore, following the setup in DeepSEM, we combine an L1 sparse loss term that regulates the sparsity of the learned adjacency matrix. The final form of the objective function is to minimize the following loss function:

	Summary of % Change			Sig. Changes	
	Min.	Mean	Max.	Better	Worse
Based on EPR					
Btw. Single-run	-0.8%	9.1%	26.8%	11	0
Btw. Ensemble	-5.8%	2.7%	21.4%	6	4
Cross compare	-7.4%	1.6%	18.8%	4	4
Based on AUPRR					
Btw. Single-run	3.3%	10.1%	25.4%	14	0
Btw. Ensemble	-5.0%	3.5%	22.4%	7	2
Cross compare	-5.9%	2.3%	19.9%	5	3

Table 1. Summary of performance changes between DAZZLE and DeepSEM across 14 BEELINE benchmarks with 1,000 target genes. Significance (adjusted p value < 0.05) is calculated by paired T test with Bonferroni correction. "Cross-compare" compares a single run of DAZZLE to the DeepSEM ensemble model.

$$\begin{aligned}\text{Loss} = &- \mathbb{E}_{z\sim q_\phi(Z|X)}[\log p_\theta(X|Z)] \\ &+ \beta D_{KL}(q_\phi(Z|X)||p_\theta(Z)) \\ &+ \gamma L_{BCE}(M_{\text{DA}}, \hat{M}_{\text{DA}}) \\ &+ \alpha \sum |A|.\end{aligned} \tag{10}$$

Additionally, we made several other model design and training choices that distinguish our model from that of DeepSEM. First, we improved the stability of model by delaying the introduction of the sparse loss term by a customizable number of epochs. Second, DeepSEM is trained on all expression data while we also explored the possibility of training exclusively on the non-zero values. This approach is discussed in more detail in the Results. Next, in terms of prior estimation, DeepSEM estimates a separate latent variable while DAZZLE uses a closed-form normal distribution. The closed-form solution reduced the number of model parameters up to 70%. Finally, while DeepSEM is trained with two separate optimizers (one on the adjacency matrix and the other on the rest of the neural networks) in an alternating manner, DAZZLE is trained using a single optimizer with different learning rates. This improvement helps DAZZLE stay modular, so it can be integrated with other network components more easily in the future.

After the DAZZLE model converges, the adjacency matrix can be extracted from the model and converted to an adjacency list. This list is then sorted by the absolute value of the edge weights and the highest valued edges are selected for evaluation.

3 Experiments and results

In this section, we start with a comparison of the performance of DAZZLE under this set of default parameters to the performance of DeepSEM (with comparisons to other methods appearing in the Supplement). Then we explore how much DAZZLE gains from each model component separately. Finally, we apply DAZZLE to a practical analysis problem, that of interpreting temporal changes in the Hammond dataset. Based on the performance on the BEELINE benchmarks, we suggest training DAZZLE with dropout augmentation intensity at 10% sampled from all data. We also suggest training DAZZLE on only non-zero expression counts and delaying the introduction of the sparse loss term by 5 steps.

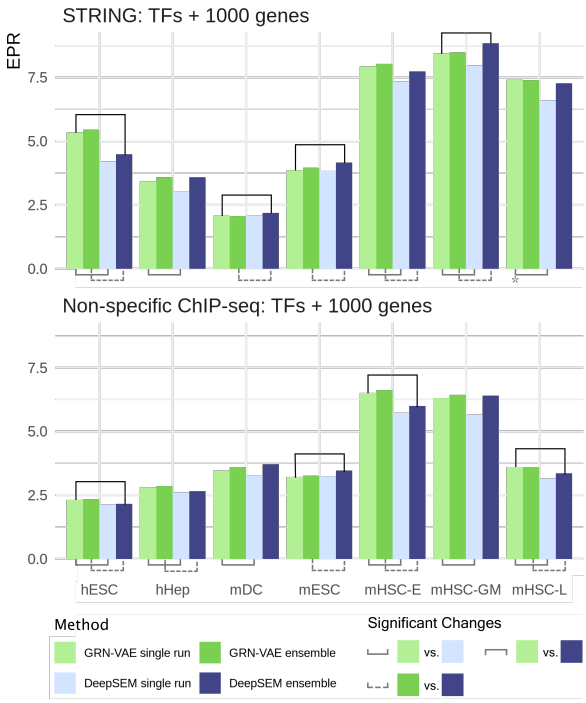

Fig. 2. Comparison of DAZZLE and DeepSEM on 14 BEELINE Benchmarks with 1,000 target genes based on EPR scores. Significant changes (adjust p value < 0.05) are marked by horizontal brackets.

3.1 GRN Inference on BEELINE benchmark data sets

A comparison of DAZZLE and DeepSEM on the BEELINE benchmarks is illustrated in Figure 2. We evaluated both methods using single runs and ensemble models across 28 BEELINE benchmarks, which consist of 7 data sets, 2 "ground truth" networks for comparison, and 2 sizes of collections of target genes. (The comparable table for these benchmarks with just 500 target genes is included in the supplementary materials.) Each ensemble model is constructed by summing the predicted adjacency matrices from 10 repetitions of the given model. The results presented in Figure 2 show the 7 data sets and represent the average EPR score of 100 repetitions for single runs and 10 repetitions for the ensemble models. The same plots for average AUPRR scores can be found in the Supplemental materials.

In addition, we summarize the comparison between DAZZLE and DeepSEM individual and ensemble models on the BEELINE benchmarks with 1,000 target genes in Table 1. We compare the EPR/AUPRR performance of DAZZLE with that of DeepSEM and highlight the percent change. We also count the number of significant changes using paired T tests adjusted for multiple comparison.

In a single run of the model, DAZZLE demonstrates superior GRN inference capabilities compared to DeepSEM. Across all benchmarks with 1,000 target genes, DAZZLE improves the prediction EPR by an average of 9.1% (ranging from -0.8% to 26.8%). In 11 out of the 14 benchmarks, these improvements are statistically significant. In the ensemble model, the performance of the ensemble models for DAZZLE and DeepSEM is comparable. The average improvement for DAZZLE is 2.7% and the numbers of significant improvements (6) and declines (4) are similar. Interestingly, DeepSEM experiences a substantial gain by combining the results from 10 repeated runs, behavior not observed with DAZZLE. A plausible explanation is that the results from each run of DAZZLE are already consistent, due to enhanced robustness and model stability. As a result, using an ensemble to combine the 10 stable and similar outcomes does not yield significant gains.

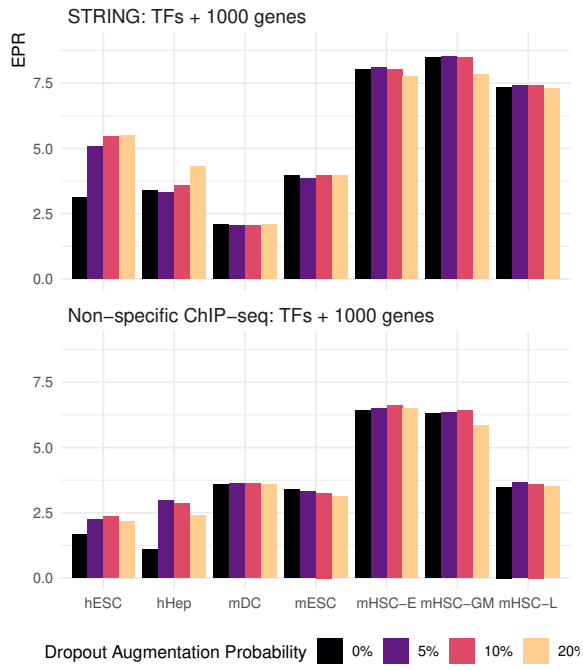

Fig. 3. Comparison of different dropout augmentation probabilities in DAZZLE.

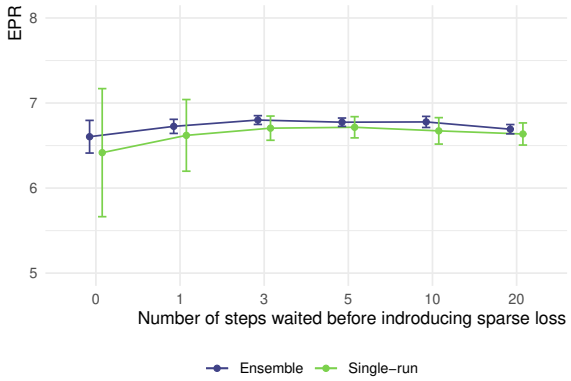

Fig. 4. Delaying the introduction of sparse loss for certain steps help reduce results variance for both single-run and ensemble models for mHSC-E with TFs + 500 genes and Non-specific ChIP-seq (a typical example; results for other benchmarks are similar).

	Summary of % Change			Sig. Changes	
	Min.	Mean	Max.	Better	Worse
Based on EPR					
Btw. Single-run	-1.6%	1.9%	6.4%	7	0
Btw. Ensemble	-1.8%	2.1%	8.5%	6	0
Based on AUPRR					
Btw. Single-run	-1.5%	1.2%	5.8%	7	4
Btw. Ensemble	-2.9%	1.2%	6.1%	6	2

Table 2. Training with only the non-zero data outperforms training with the full data. Data show reflect BEELINE benchmarks with 1,000 target genes

In fact, we even find that a single run of DAZZLE is comparable to the ensemble version of DeepSEM, as shown by the "Cross-compare" analysis in Table 1. Single-run DAZZLE outperforms ensemble DeepSEM by an average of 1.6% when evaluating EPR and 2.3% for AUPRR. The number of significant improvements and declines are similar for both the EPR and AUPRR metrics.

3.2 Impact of Dropout Augmentation

Dropout Augmentation can be applied to the model at different intensities. In the default model, 10% DA means that 10% of data are randomly chosen and set to zero in each iteration of training (whether or not they already contain zero values, so the number of data points that are changing is below 10%). Here we vary this probability from 0% to 20% while keeping all the other variables fixed to assess the impact of DA on different data sets. The results are visualized in Figure 3. For simplicity, we only display results on benchmarks with 1,000 target genes, but similar patterns are observed in benchmarks with 500 target genes.

As shown in the figure, dropout augmentation provides a significant performance lift for hESC and hHep, two human single-cell data sets in BEELINE. The increase starts to show with light augmentation at 5% and often peaks at a moderate augmentation of 10%. Heavy augmentation (20%) often start to damage the performance, but in two cases, it yields the best results. A moderate DA (10%) seems to be a good default because it either provides a gain or yields similar performance. Overall, the performance gain is inconsistent across datasets. Dropout Augmentation appears, in this case, to be more useful for the human than for mouse data sets, but more work is needed to explain this results. We also investigated the correlation of performance with the percentage of actual zero counts in the data sets, but we didn't find consistent patterns.

3.3 Impact of Training with delayed sparse loss

Both DAZZLE and DeepSEM are trained with a sparse loss term that regularizes the adjacency matrix to prevent overfitting. Experiments have shown that a large coefficient for this sparse loss is required to generate a meaningful adjacency matrix prediction in GRN inference. However, we observed that the sparse loss often destabilizes the model and traps the model in local minima. The direct outcome is that the performance of single-run DeepSEM is unstable. We therefore propose to overcome this limitation by delaying the introduction of the sparse loss term for a number of steps. Figure 4 demonstrates that delaying the addition of the sparse loss by as few as 3 to 5 epochs, we can significantly reduce the result variance in both single-run and ensemble models. A full visualization of performance for all 28 benchmarks is included in the supplementary document.

3.4 Impact of Training with Non-Zero Data

Single-cell data is well-known to be noisy. Importantly, many counts of zero in scRNA-seq data do not truly reflect a lack of expression of the corresponding transcripts. Rather, these are readings missed by the assay, especially when the true expression counts are low. 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, when we applied this idea to the original DeepSEM model, we found that it hurt performance on many benchmarks and introduced considerable variance in the results. We suspect that focusing on non-zero values may make the DeepSEM model less robust and more sensitive to random noise or other factors such as sparse loss.

With its increased model robustness and stability, DAZZLE makes training on only non-zero values a more promising choice. We find that doing so modestly improves model prediction. In Table 2, we compare the

performance of DAZZLE trained on all the data to a version trained only on non-zero data, with all other parameters fixed. Given these results, we believe that in most cases, training on non-zero data only is a good choice.

3.5 Application to a real microglial single cell data set

To test the effectiveness of running DAZZLE on actual single-cell data, we applied the method to data from mouse microglia at different developmental stages, generated by Hammond *et al.* (2019). After cell and gene filtering (detailed description of the data and preprocessing steps appears in the Methods section), the final data set includes 49,972 cells from five time points across the mouse lifespan: Embryonic (embryonic day E14, 11,262 cells and 15,673 genes), Early Postnatal (postnatal day P4/5, 13,316 cells and 15,039 genes), Late Juvenile (P30, 9,431 cells and 13,929 genes), Adulthood (P100, 8,259 cells and 13,998 genes), and Old Age (P540, 7,704 cells and 14,140 genes). At each time point, an adjacency matrix with all the input genes is calculated using DAZZLE and all the edges with absolute value larger than 0.001 are extracted and analyzed.

The GRN inference results show that gene regulation is a dynamic process that changes in different life stages. In the supplement, we list the top 10 regulated genes at each time point, ranked according to the cumulative edge weights of regulating relationships on each gene. At the earliest life stages, most of the top regulated genes are associated with cell proliferation and differentiation. For example, *Tuba1a* (Tubulin alpha 1a) encodes proteins in microtubules, which form the mitotic spindle for cell division and motility structures that move cells to their correct positions. At later ages, we see more regulation of immune response genes. Many of these genes, such as *Tmem176B* (transmembrane protein 176B), *H2.D1* (histocompatibility 2, D region locus 1), and *PISD* (phosphatidylserine decarboxylase), encode key proteins, receptors, and enzymes of microglial immune response.

In Figure 5, shows a closer view of two specific genes of interest. (Similar visualizations for other genes are available on the project website.) First, since *Tmem119* (transmembrane protein 119) is often used as a biomarker to differentiate microglia from other immune cells in the brain (Bennett *et al.*, 2016), we choose it as the center of the local network to analyze. As shown in Figure 5, *Tmem119* is only heavily active after the late juvenile stage. Among the top 10 predicted regulators, *C1qc*, *Cd81*, *Cx3cr1*, *Hexb*, *Lgmn*, *P2ry12*, and *Selplg* are commonly considered as part of the microglia transcriptional "signature," as they are generally expressed at low levels in other immune cells (Masuda *et al.*, 2020; Holtman *et al.*, 2015; Boche and Gordon, 2022; Schwabenland *et al.*, 2021; Pettas *et al.*, 2022). In addition, *Csf1r* has been recently identified as a regulator of pathogenesis for microglia and macrophages (Hagan *et al.*, 2020). It is reasonable to hypothesize that this surrounding local neighborhood includes much of the core functionality of healthy microglia.

Apoe (apolipoprotein E) is another well-studied gene that encodes a protein playing a central role in lipid metabolism, neurobiology, and neurodegenerative diseases. Unlike *Tmem119*, *Apoe* is highly regulated at embryonic day E14. As the mice mature, the relative impact of *Apoe* drops, but it increases again in old age. The list of top predicted links by DAZZLE is consistent with recent studies showing the variety of *Apoe*'s roles in many cellular activities. For example, at E14, the top three genes regulating *Apoe* are *Ftl1* (ferritin light polypeptide 1), *Tmsb4x* (thymosin beta 4 X-linked), and *Itm2b* (integral membrane protein 2B). The connection between *Apoe* and *Ftl1* is not well established, but a recent study by Ma *et al.*, 2021 shows that *Apoe* deficiency leads to increased iron levels. Another study by Kenkhuis *et al.*, 2021 suggests iron loading is a prominent marker of activated microglia in Alzheimer's disease patients. Further, both genes are located on mouse chromosome 7. This evidence suggests a plausible regulatory connection between *Ftl1* and *Apoe* and further reflects the important role of iron in early brain development.

This image illustrates DAZZLE's predicted regulation patterns for some typical well expressed genes. However, we have found that DAZZLE's regulatory predictions make sense even for genes whose overall expression levels are low. (See, e.g., *Ifit3* (Interferon Induced Protein With Tetratricopeptide Repeats 3) on the project web site.)

Another gene worth mentioning is *Malat1* (Metastasis Associated Lung Adenocarcinoma Transcript 1), which appears in our predicted networks as one of the top regulators for many microglia core genes, including (Figure 5) *Tmem119*, *Apoe*, and *H2.D1*. As a long non-coding RNA (lncRNA), *Malat1* has been identified in many pathological processes with immunological components, including (Amodio *et al.*, 2018) and diabetes (Gordon *et al.*, 2018). It has also been identified as a key regulator in the microglial inflammatory response (Zhou *et al.*, 2018; Cai *et al.*, 2020) but beyond that its function in microglia is mostly unexplored.

Overall, our analysis of this data set confirms that DAZZLE can handle typical real-world single-cell data with over ten thousand genes and thousands of cells. On examination, of the predicted networks appear generally consistent with current research on gene regulation in microglia. Beyond previously identified links, DAZZLE also suggests novel yet plausible links that may be confirmed through future experiments.

3.6 Runtime Analysis

As described in the Methods section, DAZZLE parameterizes the entire adjacency matrix. This design makes time complexity approximately quadratic in the number of genes and linear in the number of cells. With respect to clock time, DAZZLE finishes inference for one the the BEELINE benchmarks, with 758 cells and 910 genes, in an average of 16 seconds on a machine with an Nvidia A100 card. For one time point of the much larger Hammond data set, with 7,704 cells and 14,140 genes, inference takes approximately 5.25 hours.

4 Discussion

In this study, we tackle the dropout problem in real-world single-cell data paradoxically by adding more dropout. To our knowledge, this is the first time a noise-injection approach has been applied for this purpose. Our experiments show that at least for the GRN inference task, DA improves model performance.

In terms of GRN inference, our proposed method DAZZLE not only stabilizes the predictions but also produces better predictions. On the BEELINE benchmarks, a single-run of DAZZLE yields comparable results to an ensemble containing of 10 repeated trials with the previously most-accurate method.

Finally, our experiment with the microglia dataset shows that DAZZLE has the capacity to run on large single-cell data with minimal gene filtration. The predicted networks are consistent with current understanding, and include plausible novel links. These novel links could serve as good candidates for future investigations of key regulatory relationships.

One major limitation of the current model architecture is that the space complexity of this model also scales quadratically. With 15,000 genes, the model requires 30 Gb GPU memory, which still fits in a single modern GPU. However, for even larger use cases, the method may require multiple GPUs. Another limitation is that the current version is designed to be applied to each individual dataset (time point or cell cluster). Thus, it does not develop a universal understanding of gene interactions. How to lift these restrictions, how to learn these connections in a more efficient way, and how to interpret the inferred networks intuitively are the main questions to consider in future work.

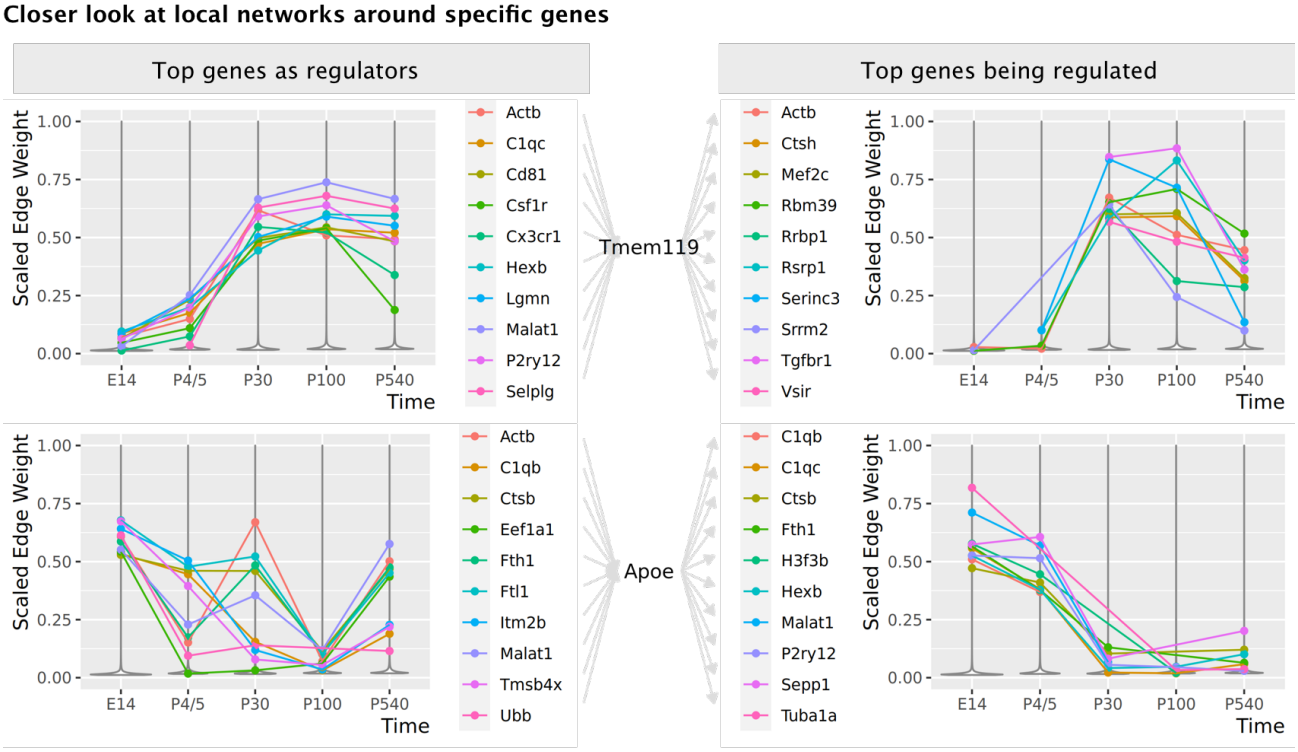

Fig. 5. Predicted local networks for Tmem119 and Apoe in mouse microglia, inferred by DAZZLE from the Hammond data. Each plot consists of two panels, where the left shows top genes that regulate the indicated gene, and the right shows top genes that it regulates. Here edge weights are scaled between 0 and 1 using the maximum weight at each time point. The backgrounds are violin plots of the scaled weights. Top genes are selected according to the maximum weight at all time points.

Acknowledgements

We thank Liping Liu, Rebecca Batorsky, and Teresa Przytycka for their thoughtful comments. We also appreciate Hantao Shu, Jianyang Zeng, and Jianzhu Ma for sharing their data and code from the original DeepSEM paper.

References

Aibar, S. *et al.* (2017). Scenic: single-cell regulatory network inference and clustering. *Nature methods*, **14**(11), 1083–1086.

Amodio, M. *et al.* (2019). Exploring single-cell data with deep multitasking neural networks. *Nature methods*, **16**(11), 1139–1145.

Amodio, N. *et al.* (2018). Malat1: a druggable long non-coding rna for targeted anti-cancer approaches. *Journal of hematology & oncology*, **11**, 1–19.

Bennett, M. L. *et al.* (2016). New tools for studying microglia in the mouse and human cns. *Proceedings of the National Academy of Sciences*, **113**(12), E1738–E1746.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*, **7**(1), 108–116.

Boche, D. and Gordon, M. N. (2022). Diversity of transcriptomic microglial phenotypes in aging and alzheimer's disease. *Alzheimer's & Dementia*, **18**(2), 360–376.

Cai, L.-J. *et al.* (2020). Lncrna malat1 facilitates inflammasome activation via epigenetic suppression of nrf2 in parkinson's disease. *Molecular brain*, **13**(1), 1–15.

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

Chen, M. and Zhou, X. (2018). Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, **19**(1), 1–15.

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

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.

ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), 57.

Eraslan, G. *et al.* (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, **10**(1), 390.

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.

Gordon, A. D. *et al.* (2018). Malat1: a regulator of inflammatory cytokines in diabetic complications. *Endocrinology, diabetes & metabolism*, **1**(2), e00010.

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.

Hagan, N. *et al.* (2020). Csf1r signaling is a regulator of pathogenesis in progressive ms. *Cell Death & Disease*, **11**(10), 904.

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.

Hinton, G. E. *et al.* (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Holtman, I. R. *et al.* (2015). Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. *Acta neuropathologica communications*, **3**, 1–18.

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.

Kenkhuis, B. *et al.* (2021). Iron loading is a prominent feature of activated microglia in alzheimer's disease patients. *Acta neuropathologica communications*, **9**(1), 1–15.

Kim, S. (2015). ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, **22**(6), 665.

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

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

Lotfollahi, M. *et al.* (2019). scgen predicts single-cell perturbation responses. *Nature methods*, **16**(8), 715–721.

Ma, J. *et al.* (2021). Apolipoprotein e deficiency induces a progressive increase in tissue iron contents with age in mice. *Redox Biology*, **40**, 101865.

Masuda, T. *et al.* (2020). Novel hexb-based tools for studying microglia in the cns. *Nature immunology*, **21**(7), 802–815.

Matsumoto, H. *et al.* (2017). Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, **33**(15), 2314–2321.

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

O'Leary, N. A. *et al.* (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**(D1), D733–D745.

Papili Gao, N. *et al.* (2018). Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, **34**(2), 258–266.

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

Pettas, S. *et al.* (2022). Profiling microglia through single-cell rna sequencing over the course of development, aging, and disease. *Cells*, **11**(15), 2383.

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

Schlauch, D. *et al.* (2017). Estimating gene regulatory networks with pandar. *Bioinformatics*, **33**(14), 2232–2234.

Schwabenland, M. *et al.* (2021). Analyzing microglial phenotypes across neuropathologies: a practical guide. *Acta neuropathologica*, pages 1–14.

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

Specht, A. T. and Li, J. (2017). Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics*, **33**(5), 764–766.

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, Y. *et al.* (2022). Netrex-cf integrates incomplete transcription factor data with gene expression to reconstruct gene regulatory networks. *Communications Biology*, **5**(1), 1282.

Xu, H. *et al.* (2013). Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database*, **2013**.

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

Zhang, S. *et al.* (2023). Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*, **14**(1), 3064.

Zhou, H.-J. *et al.* (2018). Long noncoding rna malat1 contributes to inflammatory response of microglia following spinal cord injury via the modulation of a mir-199b/ikkβ/nf-κb signaling pathway. *American Journal of Physiology-Cell Physiology*, **315**(1), C52–C61.

Zou, Z. *et al.* (2022). Chip-atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating chip-seq, atac-seq and bisulfite-seq data. *Nucleic acids research*, **50**(W1), W175–W182.