Improved gene regulatory network inference from single cell data with dropout augmentation

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

1 Department of Computer Science, Tufts University, Medford, MA, United States

* donna.slonim@tufts.edu, hao.zhu@tufts.edu

Abstract

A major challenge in working with single-cell RNA sequencing data is the prevalence of "dropout," when some transcripts' expression values are erroneously not captured. Addressing this issue, which produces zero-inflated count data, is crucial for many downstream data analyses including the inference of gene regulatory networks (GRNs). In this paper, we introduce two novel contributions. First, we propose Dropout Augmentation (DA), a simple but effective model regularization method to address the zero inflation problem in single-cell data by augmenting the data with synthetic dropout events. DA offers a new perspective to solve the "dropout" problem beyond imputation. Second, we present DAZZLE, a stabilized and robust version of the autoencoder-based structure equation model for GRN inference using the DA concept. Benchmark experiments illustrate the improved performance and increased stability of the proposed DAZZLE model over existing approaches. 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. The improved robustness and stability of DAZZLE make it a practical and valuable addition to the toolkit for GRN inference from single-cell data. Finally, we propose that Dropout Augmentation may have wider applications beyond the GRN-inference problem. Project website: https://bcb.cs.tufts.edu/DAZZLE.

Author summary

The prevalence of false zeros in single-cell data, or "dropout," affects many downstream analyses. A common approach is to eliminate these zeros through data imputation. We propose an alternative solution that focuses on regularizing the model and increasing model robustness against dropout noise. Counter-intuitively, this is done by augmenting the input data with a small number of zeros to simulate additional dropout noise. Validation is performed on the task of gene regulatory network inference. Our proposed model, DAZZLE, which uses the dropout augmentation idea, shows improved performance and robustness.

1 Introduction

Gene Regulatory Network (GRN) inference from expression data offers a contextual model of the interactions between genes *in vivo*. [1–3]. Understanding these interactions is crucial to gain insight into the development, pathology, and key points of regulation that may be amenable to therapeutic intervention [4].

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) [5]. 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. However, opportunities come with challenges. A recent benchmark paper on GRN inference summarized the major issues in single-cell data that cause challenges for GRN inference: cellular diversity, inter-cell variation in sequencing depth, cell-cycle problems, and sparsity due to dropout [6].

Despite these challenges, many methods have been proposed for context-specific GRN inference from single-cell RNA-sequencing data alone. Among established methods, GENIE3 [7] and GRNBoost2 [8] are tree-based approaches, initially proposed for bulk data, that have been found to work well on single-cell data without modification. LEAP [9] estimates pseudotime to infer gene co-expression over several lagged windows, suggesting that the lags can be used to infer regulatory relationships. SCODE [10] and SINGE [11] apply a similar pseudotime idea, combined with ordinary differential equations (ODEs) and Granger causality ensembles, to model the results. Other methods infer GRNs by integrating transcriptomic and other data sources. For example, SCENIC [12] starts by identifying gene co-expression modules using GENIE3/GRNBoost2, followed by identifying key transcription factors (TFs) that regulate these modules or regulons. scMTNI [13] studies GRNs in different cell clusters using a multi-task learning framework. GRNUlar [14] uses a recently developed unrolled algorithms to infer undirected GRNs from single-cell data by incorporating TF information. NetREX-CF [15] performs optimizations based on prior GRN networks and uses collaborative filtering to address the incompleteness of prior data. PANDA [16] further optimizes prior GRN networks using massage passing. However, single cell data alone is much more widely available and accessible in specific contexts than integrated multi-omic data sets.

The application of neural networks (NNs) in the analysis of single-cell data has advanced rapidly in the last couple of years. One of the leading NN-based GRN inference methods, DeepSEM, [17] parameterizes the adjacency matrix and uses a variational autoencoder (VAE) [18] architecture optimized on reconstruction error. In fact, on the BEELINE benchmarks [6] where the "right" networks are (approximately) known, DeepSEM reports better performance than other methods and runs significantly faster than most.

However, as shown later in this paper, one of the issues with DeepSEM is that as training continues, the quality of the inferred networks may degrade quickly. A possible explanation is that soon after the model converges, it may begin to over-fit the dropout noise in the data.

Single-cell data is often characterized by an excessive number of zero expression counts, referred to as "zero-inflation." For example, in the nine data sets examined in [19], 57 to 92 percent of the observed counts are zeros. Among these zero values, "dropout" describes the situation when transcripts, often those with low or moderate expression in a cell, are not counted by the sequencing technology. Later droplet-based protocols, such as inDrops [20] and 10X Genomics Chromium [21], helped improved detection rates. However, the "dropout" problem still persists, as even recent methods have relatively low sensitivity [22, 23].

Therefore, there has been research into data imputation methods for use in single-cell analysis. Several methods have been proposed to identify and replace missing data with imputed values [22–26]. Yet many of these methods depend on restrictive assumptions, and some require additional information, such as GRNs or bulk transcriptomic data.

In this paper, we introduce two novel contributions to the fields of single cell analysis and GRN inference. First, we propose "dropout augmentation" (DA), a novel approach

to solve the zero-inflation problem by augmenting the data with a small amount of simulated dropout noise. We found that this idea, while seemingly counter-intuitive, can effectively regularize models so that they remain robust against dropout noise.

It has long been known that by adding noise to the input data during 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 [27]. Hinton further introduced the idea of using random "dropout" on either input or model parameters to improve training performance [28]. Thus, the theoretical foundations of DA are also solid.

Our second contribution is the DAZZLE model, or *Dropout Augmentation for Zero-inflated Learning Enhancement*.DAZZLE uses the same VAE-based GRN learning framework introduced by DeepSEM and DAG-GNN [17, 29], but it employs dropout augmentation and several other model modifications. These include a new method optimizing the adjacency matrix sparsity control strategy, a simplified model structure, and a closed-formed prior. Compared to DeepSEM, DAZZLE shows better model stability and robustness in our benchmark experiments. We further illustrate how DAZZLE's network inference facilitates interpreting typical-sized data sets efficiently, in this case explaining microglial expression dynamics across the mouse lifespan.

2 Results

2.1 Dropout augmentation and the DAZZLE model

GRN inference in DAZZLE is based on the structure equation model (SEM) framework previously employed by DAG-GNN and DeepSEM [17, 29]. The adjacency matrix A is parameterized and used in both sides of an autoencoder, as shown in Figure 1. The model input is simply a single-cell gene expression matrix, where the rows correspond to cells and the columns to genes. The model is trained to reconstruct the input while the weights of the trained adjacency matrix are retrieved as a by-product of training. Since ground truth networks are never available to the training model, this type of SEM model should be considered an unsupervised learning method for GRN inference. We include a detailed explanation of why the learned adjacency matrix A represents the underlying GRN, as well as other methodological details for DA and DAZZLE, in the Methods section.

One unique design aspect that differentiates DAZZLE from DeepSEM is the use of DA, as shown in Figure 1. As a model regularization method, DA can be applied to any model design that is continuously optimized. At each training iteration, we introduce a small amount of simulated dropout noise by sampling a proportion of the expression values and setting them to zeros. With multiple training iterations, the model is exposed to multiple versions of the same data with slightly different batches of dropout noise. As a result, it is less likely to over-fit any particular batch.

DAZZLE also includes a noise classifier to predict the chance that each zero is an augmented dropout value; this classifier is trained together with the autoencoder. Since we generate the locations of the augmented dropout, we can confidently use them for training. The purpose of this classifier is to move the values that are more likely to be dropout noise to a similar region in the latent space Z', so that the decoder will learn to put less weight on them when reconstructing the input data.

We made several additional model design and training choices that further distinguish our model from that of DeepSEM. First, we improved the stability of the model by delaying the introduction of the sparse loss term by a customizeable number of epochs. Another difference is that, to estimate the prior, DeepSEM estimates a separate latent variable while DAZZLE uses a closed-form Normal distribution. These

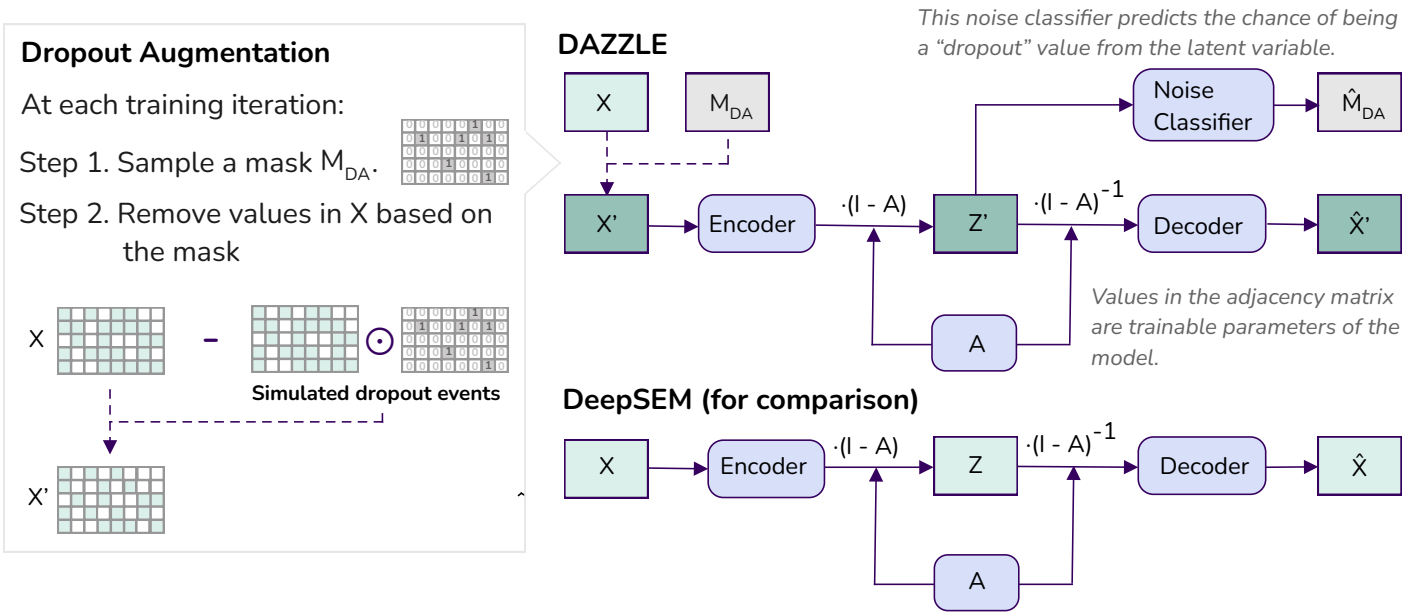

Fig 1. One of the major differences between DAZZLE and DeepSEM is the use of Dropout Augmentation. Dropout augmentation regularizes model training by simulating small amounts of random dropout at each training iteration such that the model is protected against the negative impact of dropout noise. Rounded boxes are trainable model parameters.

changes lead to reduced model sizes and computational time, especially for large networks. Finally, while DeepSEM is trained with two separate optimizers in an alternating manner (one on the adjacency matrix and the other on the rest of the neural networks), 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.

2.2 DAZZLE improves GRN inference on BEELINE benchmarks

We performed a benchmark comparison of DAZZLE to DeepSEM [17], GENIE3 [7], and GRNBoost2 [8] on the BEELINE single cell benchmark, which includes seven datasets (two from human and five from mouse) and three sets of ground truth networks. The details of the BEELINE benchmarks are described in the Methods section. We chose to compare to GENIE3 and GRNBoost2 as representatives of non-deep-learning methods because these two decision-tree-based methods are among the most widely used GRN inference algorithms. Recent benchmarks on single cell data [30], [6] also showed that these methods generally outperformed SCODE [10], ppcor [31], and SINCERITIES [32]. We also included DeepSEM, the model that inspired this work and the previous front-runner.

Note that since a single run of DeepSEM is not stable (a point that is discussed further in later Results sections). Thus DeepSEM was proposed as an ensemble algorithm, combining results from a set of 10 runs. Here, we include both single runs (1x) and the ensemble model (10x) in the comparison for both DAZZLE and DeepSEM. For fair comparison, both DAZZLE and DeepSEM use the same size neural networks, and both are trained for 120 iterations as suggested by the DeepSEM paper. The hyperparameter settings are identical except for the changes discussed in section 2.1.

The main benchmark results are provided in Table 1. The numbers reported are the

Table 1. DAZZLE shows improved GRN inference capacity on BEELINE benchmarks

	hESC	hHep	mDC	mESC	mHSC-E	mHSC-GM	mHSC-L
# of Genes	1410	1448	1321	1620	1204	1132	692
# of Cells	758	425	383	421	1071	889	847
Ground Truth: STRING							
# of True Edges	5,149	9,000	5,898	8,479	1,826	1,311	154
GENIE3	1.98 (0.01)	*1.86 (0.01)*	*1.72 (0.01)*	2.05 (0.01)	4.19 (0.05)	6.27 (0.04)	7.02 (0.08)
GRNBoost2	1.67 (0.01)	1.50 (0.01)	1.43 (0.01)	1.88 (0.02)	3.65 (0.05)	5.16 (0.11)	7.07 (0.08)
DeepSEM-1x	2.00 (0.06)	1.66 (0.04)	1.55 (0.06)	2.20 (0.02)	5.27 (0.16)	5.93 (0.38)	7.28 (0.60)
DeepSEM-10x	2.10 (0.03)	1.82 (0.05)	1.68 (0.02)	2.33 (0.03)	5.68 (0.11)	*6.91 (0.12)*	7.47 (0.14)
DAZZLE-1x	2.39 (0.05)	1.80 (0.02)	1.58 (0.03)	2.27 (0.04)	5.87 (0.21)	6.56 (0.10)	7.50 (0.07)
DAZZLE-10x	*2.44 (0.02)*	1.82 (0.01)	1.62 (0.01)	*2.34 (0.03)*	*6.07 (0.03)*	6.63 (0.04)	*7.53 (0.05)*
Ground Truth: Non-celltype-specific ChIP-Seq							
# of True Edges	4,597	5,335	3,918	8,030	1,960	1,358	317
GENIE3	0.97 (0.00)	1.01 (0.01)	1.56 (0.01)	1.65 (0.01)	2.54 (0.02)	*3.56 (0.03)*	2.78 (0.04)
GRNBoost2	1.01 (0.01)	1.09 (0.01)	1.36 (0.02)	1.49 (0.02)	2.39 (0.04)	2.84 (0.03)	2.45 (0.05)
DeepSEM-1x	1.22 (0.05)	1.29 (0.08)	1.73 (0.07)	1.58 (0.05)	2.94 (0.32)	3.27 (0.15)	2.50 (0.56)
DeepSEM-10x	1.24 (0.01)	1.41 (0.03)	*1.93 (0.02)*	*1.66 (0.02)*	3.18 (0.06)	3.48 (0.12)	3.01 (0.20)
DAZZLE-1x	1.27 (0.05)	*1.44 (0.04)*	1.85 (0.04)	1.62 (0.05)	3.30 (0.05)	3.45 (0.06)	3.00 (0.12)
DAZZLE-10x	*1.29 (0.01)*	1.44 (0.01)	1.89 (0.02)	1.64 (0.01)	*3.35 (0.02)*	3.51 (0.03)	*3.21 (0.11)*
Ground Truth: Celltype-specific ChIP-Seq							
# of True Edges	7,050	15,410	1,193	42,795	21,975	14,135	5,180
GENIE3	0.96 (0.00)	*1.10 (0.00)*	1.01 (0.01)	1.06 (0.00)	0.99 (0.00)	0.98 (0.00)	1.05 (0.00)
GRNBoost2	1.01 (0.00)	1.02 (0.00)	1.00 (0.01)	1.02 (0.00)	0.99 (0.00)	0.99 (0.00)	1.01 (0.00)
DeepSEM-1x	1.06 (0.02)	1.03 (0.01)	1.04 (0.03)	1.03 (0.01)	1.01 (0.00)	1.02 (0.01)	1.06 (0.01)
DeepSEM-10x	1.07 (0.01)	1.03 (0.00)	*1.05 (0.01)*	1.03 (0.00)	1.01 (0.00)	1.02 (0.00)	1.06 (0.00)
DAZZLE-1x	1.10 (0.02)	1.02 (0.01)	1.03 (0.03)	1.03 (0.01)	1.01 (0.01)	1.01 (0.01)	1.08 (0.01)
DAZZLE-10x	*1.10 (0.01)*	1.03 (0.00)	1.05 (0.01)	*1.03 (0.01)*	*1.01 (0.00)*	1.08 (0.00)	*1.08 (0.00)*

Metric: AUPRC Ratio; Number of target genes: 1000.
Numbers reported are mean and std. of AUPRC Ratio compared with random guess over 10 runs. Higher ratios indicate better performance. Here, italicized text in dark shading indicate the best algorithms and lightly shaded cells indicate the 2nd best algorithms. In this experiment, we ran each algorithm on all available data with the default settings (for cell-specific ones, we followed the suggestions from BEELINE and used very small L1 regulation on the adjacency matrix). The 10x models ensemble inferred networks from 10 runs and the 1x models are simply single-run models.

average area under the precision recall curve (AUPRC) ratios over 10 runs, where higher values represent better performance. In Table 1, the evaluation is done separately for the STRING network, the Non-celltype-specific ChIP-Seq network, and the Celltype-specific ChiP-Seq network. Note that for celltype-specific ChiP-Seq, we followed the recommendations of the DeepSEM paper and applied a very small L1 sparsity regulation to the adjacency matrix for better performance.

Among all the evaluations on the STRING and Non-celltype specific benchmarks, the ensemble version of DAZZLE (DAZZLE-10x) has the best performance of all methods in over half the cases, and in all the other cases it ranks either 2nd or 3rd, with a result that is within 6% of the top score. The DAZZLE-10x results also typically have lower variance compared to the DeepSEM-10x results, illustrating the stability of the DAZZLE model. Further, across nearly all benchmarks, a single run of DAZZLE, DAZZLE-1x, has a higher AUPRC ratio than the comparable single-run DeepSEM-1x. Yet, DAZZLE-1x is typically also more stable. Some DeepSEM-1x results in Table 1 have substantial variance, but this unwanted model behavior is not shown by DAZZLE-1x; the reasons for this are explained in Section 2.4. We also see similar

findings when we use Early Precision Ratio (EPR) as an evaluation metric (Supplement Table S1).

For the cell-type specific data, the highest observed AUPRC ratio for any data set or method is 1.10, and the highest observed EPR is 1.20 (just slightly better than random performance), suggesting that for all methods, reproducing these cell-specific networks is nearly impossible. A possible explanation of this effect appears in Supplement Figure S1, which shows that the results are highly variable across training iterations. We confirmed that using very low L1 regularization yields better performance on this ground truth network; the reason for this is still unclear.

In summary, under similar training conditions, DAZZLE outperforms DeepSEM and demonstrates more stable performance. We found that while a single run of DeepSEM can produce unstable results, the ensemble strategy proposed in the DeepSEM paper significantly enhances its performance and reduces variance. However, this improvement comes at the cost of a tenfold increase in execution time. For DAZZLE, the ensemble strategy also yields excellent results, making DAZZLE-10x the best performer in our comparison. Note however that the performance of DAZZLE-1x itself approaches that of an ensemble method, in essence because it functions as an ensemble method, as described in the Methods section below. For datasets with a large number of genes and cells, a single pass of DeepSEM or DAZZLE can take several hours to compute, even on modern GPUs. In such cases, a single pass of DAZZLE offers a sufficiently-good solution at a more reasonable cost.

2.3 Dropout Augmentation contributes to model robustness

In this section, we discuss the effects of dropout augmentation alone on GRN inference. Supplement Figure S1 shows the quality of the inferred networks from DeepSEM-1x as a function of the number of training steps. Specifically, it shows that the quality of the inferred networks from DeepSEM tends to drop quickly if the model was over-trained, suggesting that the model may be overfitting to unwanted patterns in the training data. One possible solution is to stop training early, as suggested in the DeepSEM paper, which ends model training at 120 iterations. The tricky part is that the point of peak performance is very difficult to capture and may depend highly on the number of genes and cells in the dataset. In practice, when ground-truth networks are not available, it is very difficult to identify a good convergence point for the model. It would be ideal if inferred GRN accuracy remained robust, or at least if it dropped slowly and consistently after the performance peak, so that the cost of picking a sub-optimal stopping point would be low. We believe that dropout augmentation is potentially an effective solution to this problem.

For fairness, all the comparisons shown in Figure 2 were performed on DAZZLE. We varied only the percentage of augmented dropout values while keeping all the other hyper-parameters the same. Without DA (thick black lines, representing a DA probability of 0%), a common pattern is that the AUPRC Ratios drop after the performance peaks, similar to what we observed with the DeepSEM method (Supplement Figure S1). However, when we train the model with some amount of DA, in most cases the AUPRC Ratios either stay flat or decrease at a slower pace. For some data sets, such as hESC and hHep, DA also improves accuracy significantly. However, we observed that different datasets and ground truth networks seem to require different optimal levels of dropout augmentation. For hESC and hHep on the STRING network, 20% dropout augmentation yields the best AUPRC scores, but that amount of DA is too high for mHSC-E, mHSC-GM, and mHSC-L since it either slows down the model's convergence or leads to lower accuracy. After reviewing all cases, we recommend using 10% DA probability as the default for DAZZLE.

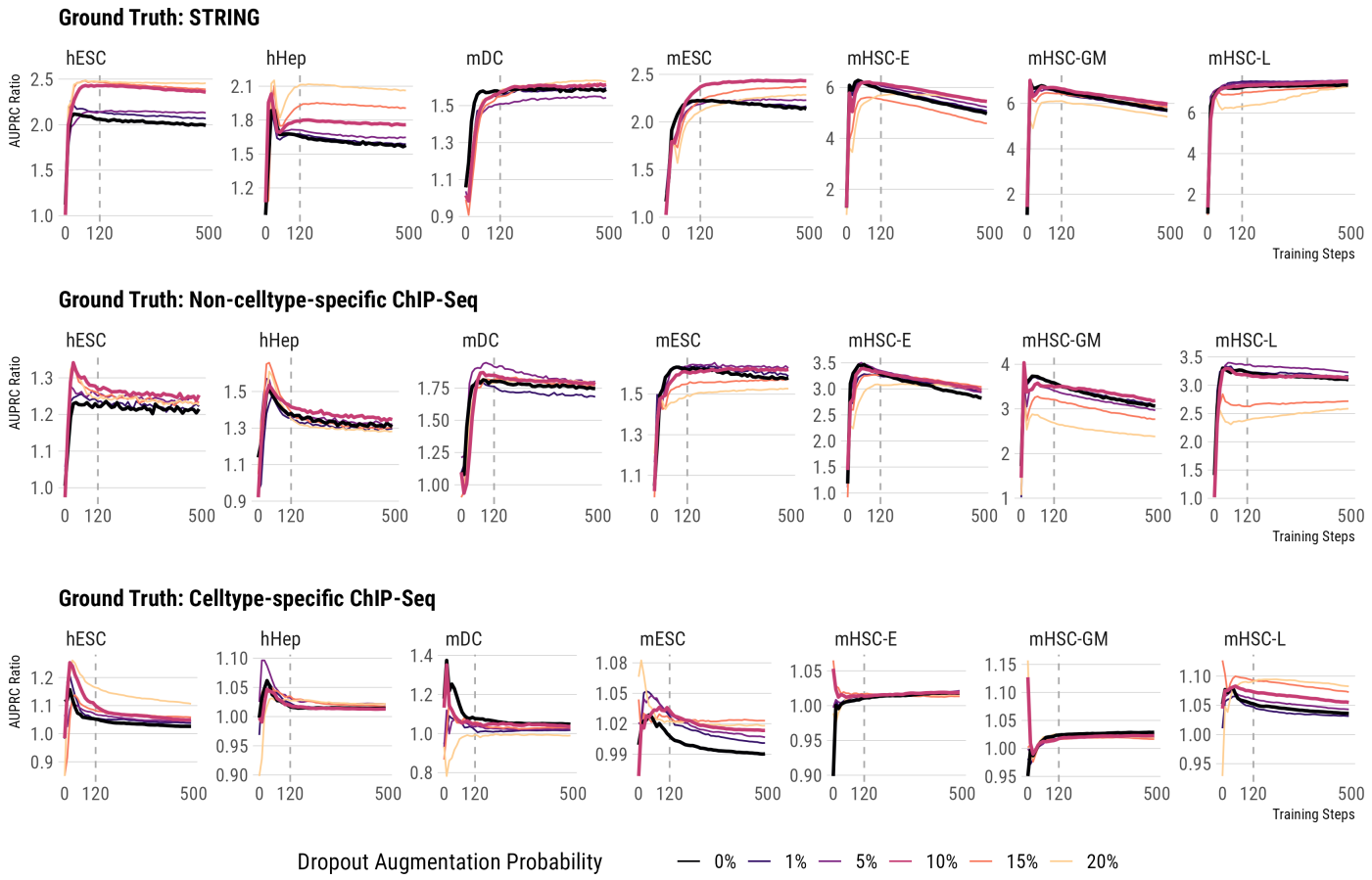

Fig 2. Appropriate amount of augmented dropout helps maintain model robustness and may contribute to better performance. Color reflects the probability of dropout augmentation. The two thick lines represent two important conditions, 0% - no augmented dropout, and 10% - the default dropout augmentation level we recommend. Dashed lines shows the default number of training iterations used in DeepSEM and DAZZLE.

2.4 Sparsity control strategy improves DAZZLE's stability

In addition to increased robustness, DAZZLE also generates more stable GRN inference results thanks to its improved sparsity control strategy. As mentioned in Section 2.2, results from DeepSEM-1x are not stable. Figure 3 examines this issue using 100 runs of DeepSEM-1x and DAZZLE-1x on the hESC dataset evaluated using the STRING network. In Figure 3a, we show histograms of the AUPRC ratios for both methods. The accuracy of the 100 runs of DeepSEM-1x can be separated into two groups. As shown in Supplement Figure S2, other benchmark data sets produced similar results. Further investigation suggested that the L1 regularization on the adjacency matrix might be the main cause of those less ideal results.

Both DAZZLE and DeepSEM are trained with L1 regularization on the adjacency matrix. This regularization term ensures that the model doesn't add unnecessary weights to the adjacency matrix. It also ensures the singularity of the inferred adjacency matrix. Experiments have shown that a large enough coefficient for this sparse loss is required to generate a meaningful adjacency matrix prediction in GRN inference. However, as shown in Figure 3b, for DeepSEM-1x, where this L1 sparse loss is

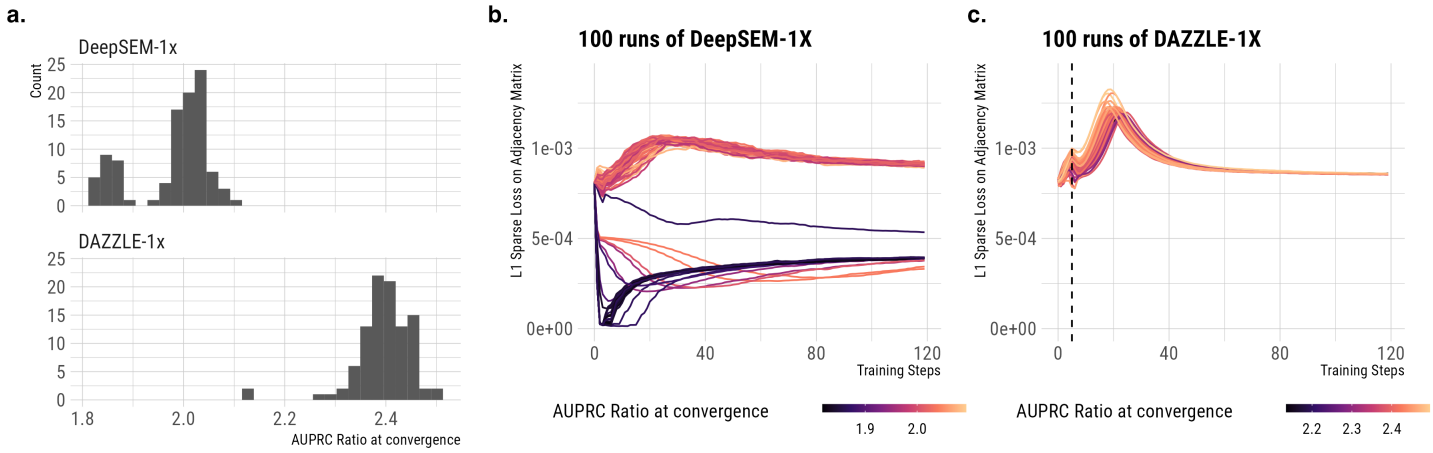

Fig 3. Comparison of 100 runs of DeepSEM-1x and DAZZLE-1x on hESC evaluated using the STRING network. In DeepSEM-1x, the prioritization of the L1 sparsity control at early stage is the main cause of unstable GRN inference performances. DAZZLE solved this issue by delaying the introduction of L1 sparse loss by 5 iterations. **a)** Histogram of AUPRC ratios. **b)** Sparse loss over time for DeepSEM-1X, colored by AUPRR at convergence. **c)** Sparse loss over time for DAZZLE-1x, colored by AUPRR at convergence.

introduced at the very beginning of the training process, the optimization of this loss may be prioritized such that the values in the adjacency matrix quickly drop to near zero within the first several training steps. Most runs that do this end up in the low performance group. In DAZZLE, we propose a simple solution to overcome this limitation by delaying the introduction of the L1 sparse loss by a few training steps. As shown in Figure 3c, after 5 steps of training without the L1 sparse loss, the values and the gradients on the adjacency matrix parameters are stabilized sufficiently so the introduction of the sparse loss is less likely to destabilize the training process. We also see similar findings on other benchmark datasets; detailed results are included in Supplement Figure S2.

2.5 Case study: Using DAZZLE to predict temporal changes in GRNs in mouse microglia

To test the effectiveness of running DAZZLE on actual single-cell data, we applied the method to published data characterizing mouse microglia at different developmental stages [33]. For each time point, DAZZLE inferred an adjacency matrix for all the input genes. All edges whose weights have an absolute value over 0.001 are extracted and analyzed.

These GRN inference results confirm that gene regulation is a dynamic process that changes across the lifespan. Figure 4a lists the top ten regulated genes at each time point, ranked by the summed 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, however, 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.

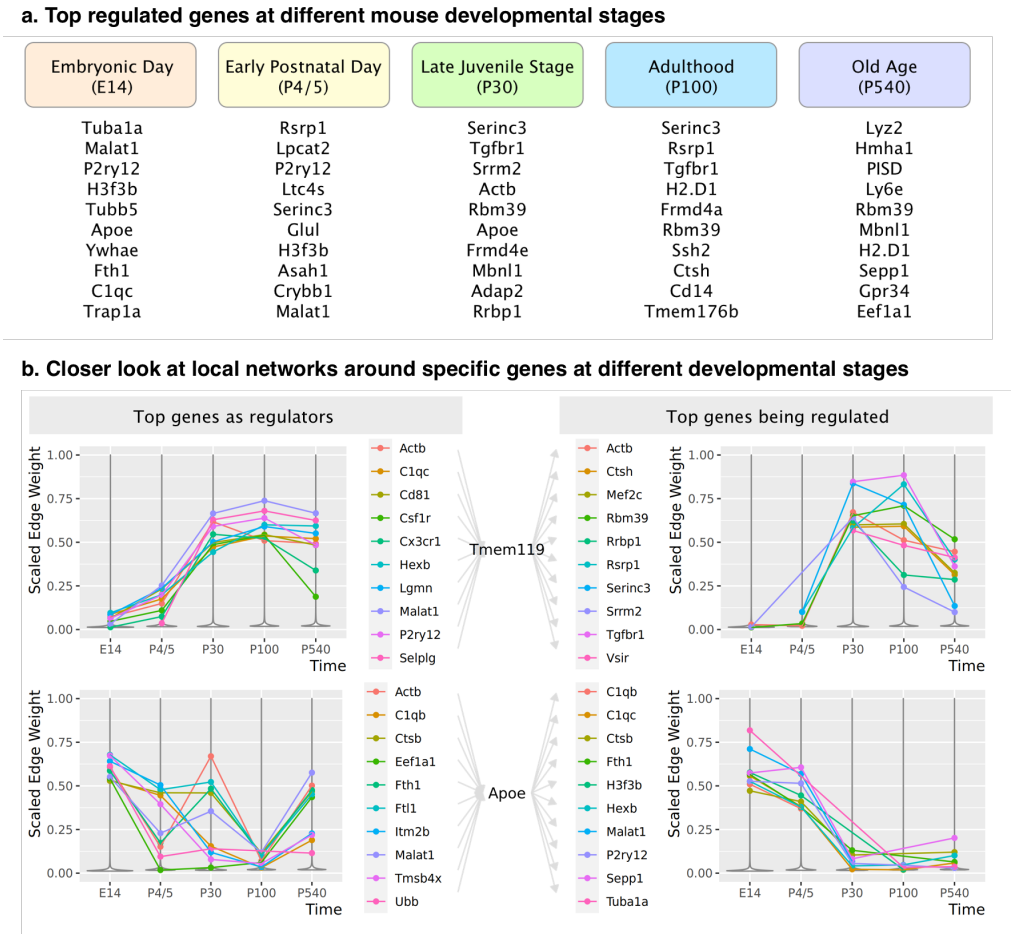

Fig 4. a. Top 10 regulated genes at each life stage in mouse microglia. b. Predicted local networks around *Tmem119* and *Apoe* in mouse microglia. Here edge weights are min-max scaled at each time point. Top genes are selected according to the maximum weight at all time points.

Figure 4b shows a closer view of two specific genes of interest. First, since *Tmem119* (transmembrane protein 119) is often used as a biomarker to differentiate microglia from other immune cells in the brain [34], we chose it as the center of the local network to analyze. Figure 4b shows that regulation of and by *Tmem119* is only heavily active starting in 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 [35–39]. In addition, *Csf1r* has been recently identified as a regulator of pathogenesis for microglia and macrophages [40]. 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 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). Although the connection between *Apoe* and *Ftl1* is not well established, a recent study [41] shows that *Apoe* deficiency leads to increased iron levels. Another study [42] suggests iron loading is a prominent marker of activated microglia in Alzheimer's disease patients. Further, both genes are located on mouse chromosome 7, about 20cM apart. This evidence suggests a plausible regulatory connection between *Ftl1* and *Apoe* and further reflects the important role of iron in early brain development.

Figure 4 illustrates DAZZLE's predicted regulation patterns for two 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., comparable images for *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 4) *Tmem119*, *Apoe*, and *H2.D1*. As a long non-coding RNA (lncRNA), *Malat1* has been identified in many pathological processes with immunological components, including several types of cancer [43] and diabetes [44]. It has also been identified as a key regulator in the microglial inflammatory response [45, 46], 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 more than 10,000 genes and thousands of cells. On examination, 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 Materials and Methods

3.1 Dropout Augmentation

Previous research suggests that zero values in single cell data include both biologically real 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 $X \in \mathbb{R}^{n \times m}$ be a gene expression matrix from a single-cell experiment, where n is the number of cells and m is the number of genes. We randomly sample a proportion (p) 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 Bernoulli 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:

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

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

In theory, the dropout augmentation algorithm could be applied to any iterative learning algorithm, such as expectation maximization, gradient boosting, or neural networks. As shown in Figure 1, M_{DA} is re-sampled at every training step, so the augmented data X' changes in every iteration. During the entire training process, the model is only exposed to the altered X' instead of X.

3.2 GRN Inference with DAZZLE

The task of GRN inference is to infer a weighted adjacency matrix $A \in \mathbb{R}^{m \times m}$ based on the expression data X. Previous methods DAG-GNN [29] and DeepSEM [17] rely on a linear additive assumption that can be written as

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

where $Z \in \mathbb{R}^{n \times m}$ 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). To embrace the idea that the observed data are noisy due to dropout, we modify this assumption and rewrite equation 3 as:

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

where Z' is defined for X' analogously to the definition of Z for X. Since dropout is so prevalent in single-cell data, we believe equation 4 describes the actual situation more accurately. 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 can be denoted by $q_\phi(Z'|X')$ and the decoder can be denoted by $p_\theta(X'|Z')$, 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, it is common to maximize the evidence lower bound (ELBO), which we write as

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

where the first term is the KL divergence and the second term can be thought of 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 can be used to infer whether a value comes from a dropout event. 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:

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

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

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

The dropout augmentation classifier can be trained either separately or together with the main model using the same optimizer. In our experiments, we add the

classification loss to the ELBO function scaled by a hyper parameter χ. Additionally, following the design of DeepSEM, we include 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:

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

3.3 Datasets Used

3.3.1 BEELINE single-cell benchmarks

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 [6]. 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 v11 [47], non cell-type specific transcriptional networks, and cell-type specific networks. The Non-specific ChIPSeq network combines links from DoRothEA [48], RegNetwork [49], and TRRUST v2 [50]. The cell-type specific networks were created by the BEELINE authors for each dataset by searching through the ENCODE [51], ChIP-Atlas [52], and ESCAPE [53] 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.

3.3.2 Hammond microglial data

To assess DAZZLE's performance in a more practical context, we use a published data set from [33] (data available from NCBI's Gene Expression Omnibus database [54] 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, each of which includes single cell data from four healthy male mice. To preprocess the data, following suggestions from [55] for the same data, we filtered out cells with fewer than 400 or more than 3,000 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, keeping only 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 gene models, mitochondrial genes, and ribosomal genes from this pool to simplify the interpretation of the resulting networks. The expression values were normalized using natural log transformation.

After this cell and gene filtering, the final data set includes 49,972 cells from five time points across the mouse lifespan: Embryonic (embryonic day E14, 15,673 genes and 11,262 cells), 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). Note that compared to the original paper [33], 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.

3.4 Evaluation metrics

In this report, we follow the recommendations from the BEELINE paper and use Area Under the Precision Recall Curve Ratio (AUPRC Ratio) and Early Precision Ratio as the main evaluation metrics. Compared to the Area Under the Receiver Operating Characteristic (AUROC), AUPRC is a better metric when there is a class imbalance between the positive and negative groups [56].

In this paper, AUPRC is approximated using its discretized form without interpolation, as shown in Equation 11:

$$\text{AUPRC} = \int_0^1 P(R)dR \approx \sum_{n=1}^{N} P_n \cdot (R_n - R_{n-1}), \tag{11}$$

where $P(R)$ is the Precision at Recall level R, N is the number of unique predictions, and R_n and P_n are the Recall and Precision scores at item n. This approximation is often referred to as average precision [57] and is the same analysis performed in the DeepSEM paper. The AUPRC Ratio is calculated by dividing the calculated AUPRC score with the theoretical AUPRC score of a random predictor. In this case, the expected precision of a random predictor is equal to the proportion of positive cases. On the Precision-Recall plot, the performance of a random predictor forms a horizontal straight line (see Figure 5D in [56]). Therefore, the AUPRC of a random predictor is simply the proportional of positive cases and we can calculate AUPRC Ratio with Equation 12.

$$\text{AUPRC Ratio} \approx \frac{\text{Positives}}{\text{Total Instances}} \cdot \sum_{n=1}^{N} P_n \cdot (R_n - R_{n-1}) \tag{12}$$

Early Precision (EP) is the fraction of true positives within the top-k candidates, where k is the number of edges in the ground truth network. Early Precision Ratio further divides that fraction with the expected Early Precision of a random predictor, which is the edge density of the ground truth network [6].

Since the number of all possible edges and the number of true edges are usually very large in actual GRNs, the values of the AUPRC and Early Precision themselves tend to be small. As claimed in the BEELINE paper, converting them to ratios make it easier to understand the performance across benchmarks.

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 for the GRN inference task, DA improves model performance.

In terms of GRN inference, our proposed method DAZZLE not only stabilizes the output 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. While we do not

have ground-truth networks suitable for direct comparison in these cases, the predicted networks are consistent with current understanding and include plausible novel links. These novel links may be 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 help us develop a universal understanding of gene interactions. How to relax these restrictions, how to learn these connections in a more efficient way, and how to interpret the inferred networks biologically are important questions to consider in future work.

Acknowledgements

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

References

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

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

3. Penfold CA, Wild DL. How to infer gene networks from expression profiles, revisited. Interface focus. 2011;1(6):857–870.

4. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Frontiers in cell and developmental biology. 2014;2:38.

5. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature protocols. 2018;13(4):599–604.

6. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature methods. 2020;17(2):147–154.

7. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PloS one. 2010;5(9):e12776.

8. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35(12):2159–2161.

9. Specht AT, Li J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics. 2017;33(5):764–766.

10. Matsumoto H, Kiryu H, Furusawa C, Ko MS, Ko SB, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33(15):2314–2321.

11. Deshpande A, Chu LF, Stewart R, Gitter A. Network inference with Granger causality ensembles on single-cell transcriptomics. Cell reports. 2022;38(6).

12. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nature methods. 2017;14(11):1083–1086.

13. Zhang S, Pyne S, Pietrzak S, Halberg S, McCalla SG, Siahpirani AF, et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. Nature Communications. 2023;14(1):3064.

14. Shrivastava H, Zhang X, Song L, Aluru S. GRNUlar: a deep learning framework for recovering Single-Cell gene regulatory networks. Journal of Computational Biology. 2022;29(1):27–44.

15. Wang Y, Lee H, Fear JM, Berger I, Oliver B, Przytycka TM. NetREX-CF integrates incomplete transcription factor data with gene expression to reconstruct gene regulatory networks. Communications Biology. 2022;5(1):1282.

16. Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing messages between biological networks to refine predicted interactions. PLoS One. 2013;8(5):e64832.

17. Shu H, Zhou J, Lian Q, Li H, Zhao D, Zeng J, et al. Modeling gene regulatory networks using neural network architectures. Nature Computational Science. 2021;1(7):491–501.

18. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013;.

19. Ghazanfar S, Bisogni AJ, Ormerod JT, Lin DM, Yang JY. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. BMC systems biology. 2016;10(5):11–24.

20. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–1201.

21. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017;8(1):14049.

22. Rao J, Zhou X, Lu Y, Zhao H, Yang Y. Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. Iscience. 2021;24(5).

23. Kang Y, Zhang H, Guan J. scINRB: single-cell gene expression imputation with network regularization and bulk RNA-seq data. Briefings in Bioinformatics. 2024;25(3):bbae148.

24. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature communications. 2018;9(1):997.

25. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome biology. 2019;20:1–14.

26. Qi Y, Han S, Tang L, Liu L. Imputation method for single-cell RNA-seq data using neural topic model. GigaScience. 2023;12:giad098.

27. Bishop CM. Training with noise is equivalent to Tikhonov regularization. Neural computation. 1995;7(1):108–116.

28. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580. 2012;.

29. Yu Y, Chen J, Gao T, Yu M. DAG-GNN: DAG structure learning with graph neural networks. In: International Conference on Machine Learning. PMLR; 2019. p. 7154–7163.

30. Kang Y, Thieffry D, Cantini L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. Frontiers in genetics. 2021;12:362.

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

32. Papili Gao N, Ud-Dean SM, Gandrillon O, Gunawan R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics. 2018;34(2):258–266.

33. Hammond TR, Dufort C, Dissing-Olesen L, Giera S, Young A, Wysoker A, et al. Single-cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. Immunity. 2019;50(1):253–271.

34. Bennett ML, Bennett FC, Liddelow SA, Ajami B, Zamanian JL, Fernhoff NB, et al. New tools for studying microglia in the mouse and human CNS. Proceedings of the National Academy of Sciences. 2016;113(12):E1738–E1746.

35. Masuda T, Amann L, Sankowski R, Staszewski O, Lenz M, d´ Errico P, et al. Novel Hexb-based tools for studying microglia in the CNS. Nature immunology. 2020;21(7):802–815.

36. Holtman IR, Raj DD, Miller JA, Schaafsma W, Yin Z, Brouwer N, et al. Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. Acta neuropathologica communications. 2015;3:1–18.

37. Boche D, Gordon MN. Diversity of transcriptomic microglial phenotypes in aging and Alzheimer's disease. Alzheimer's & Dementia. 2022;18(2):360–376.

38. Schwabenland M, Brück W, Priller J, Stadelmann C, Lassmann H, Prinz M. Analyzing microglial phenotypes across neuropathologies: a practical guide. Acta neuropathologica. 2021; p. 1–14.

39. Pettas S, Karagianni K, Kanata E, Chatziefstathiou A, Christoudia N, Xanthopoulos K, et al. Profiling Microglia through Single-Cell RNA Sequencing over the Course of Development, Aging, and Disease. Cells. 2022;11(15):2383.

40. Hagan N, Kane JL, Grover D, Woodworth L, Madore C, Saleh J, et al. CSF1R signaling is a regulator of pathogenesis in progressive MS. Cell Death & Disease. 2020;11(10):904.

41. Ma J, Qian C, Bao Y, Liu MY, Ma HM, Shen MQ, et al. Apolipoprotein E deficiency induces a progressive increase in tissue iron contents with age in mice. Redox Biology. 2021;40:101865.

42. Kenkhuis B, Somarakis A, de Haan L, Dzyubachyk O, IJsselsteijn ME, de Miranda NF, et al. Iron loading is a prominent feature of activated microglia in Alzheimer's disease patients. Acta neuropathologica communications. 2021;9(1):1–15.

43. Amodio N, Raimondi L, Juli G, Stamato MA, Caracciolo D, Tagliaferri P, et al. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. Journal of hematology & oncology. 2018;11:1–19.

44. Gordon AD, Biswas S, Feng B, Chakrabarti S. MALAT1: a regulator of inflammatory cytokines in diabetic complications. Endocrinology, diabetes & metabolism. 2018;1(2):e00010.

45. Zhou HJ, Wang LQ, Wang DB, Yu JB, Zhu Y, Xu QS, et al. 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. 2018;315(1):C52–C61.

46. Cai LJ, Tu L, Huang XM, Huang J, Qiu N, Xie GH, et al. LncRNA MALAT1 facilitates inflammasome activation via epigenetic suppression of Nrf2 in Parkinson's disease. Molecular brain. 2020;13(1):1–15.

47. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 2019;47(D1):D607–D613.

48. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome research. 2019;29(8):1363–1375.

49. Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database. 2015;2015.

50. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic acids research. 2018;46(D1):D380–D386.

51. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

52. Zou Z, Ohta T, Miura F, Oki S. 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. 2022;50(W1):W175–W182.

53. Xu H, Baroukh C, Dannenfelser R, Chen EY, Tan CM, Kou Y, et al. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database. 2013;2013.

54. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.

55. Green LA, O'Dea MR, Hoover CA, DeSantis DF, Smith CJ. The embryonic zebrafish brain is seeded by a lymphatic-dependent population of mrc1+ microglia precursors. Nature Neuroscience. 2022;25(7):849–864.

56. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3):e0118432.

57. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008. Available from: `http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html`.