MACHINE LEARNING ASSISTED GENE EXPRESSION RESTORATION

Arnob Saha Ankon
1024052111
Contact: 01781820908
ararnob9063@gmail.com

Arnab Bhattacharjee
1024052106
Contact: 01879038263
arnabbndc@gmail.com

Abstract

Gene expression is the process by which genes produce proteins that control most cellular activities. Understanding the interactions between genes and transcription factors (TFs) is important for studying cell states and detecting diseases. Existing approaches, such as GAGER [1], assume a simple linear relationship, which may not capture the true nature of gene regulation. In this work, we explore gene expression restoration using nonlinear models. We test these models on multiple scRNA-seq datasets, including Hypoplastic Left Heart Syndrome (HLHS), mouse neurodegeneration progression and *Saccharomyces cerevisiae*. Our results show that nonlinear models often perform better than linear model in reducing expression differences by using fewer TFs for gene expression restoration. For instance, the quadratic model performs consistently well across different thresholds in the mouse neurodegeneration dataset. These findings suggest that gene regulation is often nonlinear and using more flexible models can improve the identification of key TFs and genes required for gene expression restoration.

Keywords: Gene Regulatory Network (GRN), Transcription Factor (TF), Gene, Gene Expression, Gene Regulation, Nonlinearity, scRNA Sequencing

1 Introduction

A gene is a DNA sequence that codes for a functional product, such as a protein. Genes express themselves by producing proteins. Gene expression refers to the presence of proteins in a cell. Proteins are produced from genes in two stages. First, mRNA is produced from the DNA sequence, and then protein is synthesized from the mRNA. These stages are called transcription and translation, respectively. The transcription factor is a protein that regulates gene expression by controlling how much protein will be produced from a gene. Its nature can be either inhibitory or activatory.

Gene expression is essential for cellular functions, as proteins control most cellular processes. For instance, genes responsible for muscle contraction are highly expressed in muscle cells. Gene expression can be used to determine whether a cell is healthy or diseased. This gene expression depends on the cell's state, which is influenced by its previous state. We can categorize the cell's state into the source state and the target state. The changes between the source and target states occur due to the activation and repression of some of the genes present in the cell. Identification of the genes that drive the transformation from the source state to the target state is important, as this transformation enables early detection of diseases like cancer and tumors. However, identifying these genes remains an unresolved challenge.

We can represent the complex relationships between genes and transcription factors that regulate and determine cell behavior using a gene regulatory network (GRN) [2,3]. In GRNs, the nodes represent genes, and the edges represent the regulatory relationships between genes. Knowledge of the GRNs of the source and target cell states can be helpful in accurate identification of target genes. GRNs are extremely important as they help us understand the biological processes within the cell. One key application of GRNs is the comparison of healthy versus diseased tissues [7, 8]. Such studies help us identify genes that may be responsible for diseases.

Some approaches to inferring the relationship between the source state and target state include differential expression analysis [10], machine learning and deep learning-based perturbation [11] algorithmic approaches [1], or manual perturbation of specific transcription factors (TFs). In the paper GAGER (gene regulatory network assisted gene expression restoration) [1], the authors designed an algorithm to identify specific genes whose manipulation transforms the source state into the target state of the cell by comparing the GRNs of the source and target cell states. The algorithm mainly applies a series of perturbations to convert the source cell state into the target cell state. It aims to infer the differences between the regulatory edges of the GRNs of the source and target cell states. From this, it identifies TFs that potentially change the cell state from source to target.

In their work, they assume a simple linear relationship between genes, while the relationship between genes may actually be more complex. They try to perturb transcription factors (TFs) greedily to reduce the expression differences between the source and target cell states.

In the paper [4], the authors show that there are four main types of relationships between a TF and its regulated genes. A linear relationship occurs only when there is a strong correlation between a TF and its target gene. Other studies [5, 6] also present nonlinear dynamics in transcriptional regulation.

The relationship between a TF and its target gene is not always linear. Some TFs work in a cooperative manner and regulatory networks often involve feedback mechanisms. We aim to further explore gene expression restoration by assuming nonlinear relationships and identify genes or TFs required for restoration.

2 Methodology

We take the gene expression matrix and GRNs of the source and target states as input, and as output, we provide the transcription factors (TFs) to perturb and the resultant expression difference between the source and the inferred state. The overview of our methodology is as follows:

1. **Construction of the Difference Network:** We construct the difference network using the GRNs and expression matrix, as outlined in GAGER [1]. Here, we identify the nodes that have expression differences between the source and target states.

2. **Estimation of Gene Expression and Regulatory Coefficients:** We assume different nonlinear relationships between genes, rather than a simple linear relationship, to calculate the expression and regulatory coefficients of the difference network.

3. **Selection of Target Genes and Generation of GRNs:** We select relevant transcription factors and genes and update their expression values. Using the calculated coefficients, we then update the expression values of all other genes regulated by the selected TFs or genes. We then attempt to infer the GRN. Finally, we compare the inferred results with the GRN of the source cell state. As a measure, we calculate the expression difference between the source state and the inferred state.

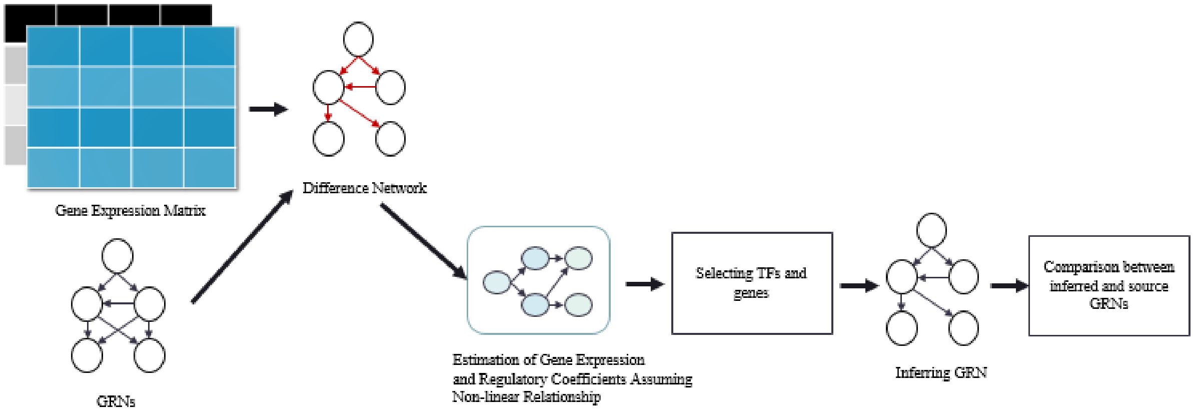

Gene Expression Matrix
Difference Network
GRNs
Estimation of Gene Expression and Regulatory Coefficients Assuming Non-linear Relationship
Selecting TFs and genes
Inferring GRN
Comparison between inferred and source GRNs

Figure 1: Overview of the methodology: steps include inferring the difference network, estimating gene expression and regulatory coefficients, selecting target genes/TFs, and comparing the inferred GRN with the target GRN.

3 Datasets

We will experiment on the following datasets:

1. **Hypoplastic Left Heart Syndrome (HLHS) scRNA-seq data:** Gene expression data from Omnibus GSE146341 (link).

2. **Progression of Neurodegeneration in Mouse Model scRNA-seq data:** Gene expression data from Omnibus GSE103334 (link).

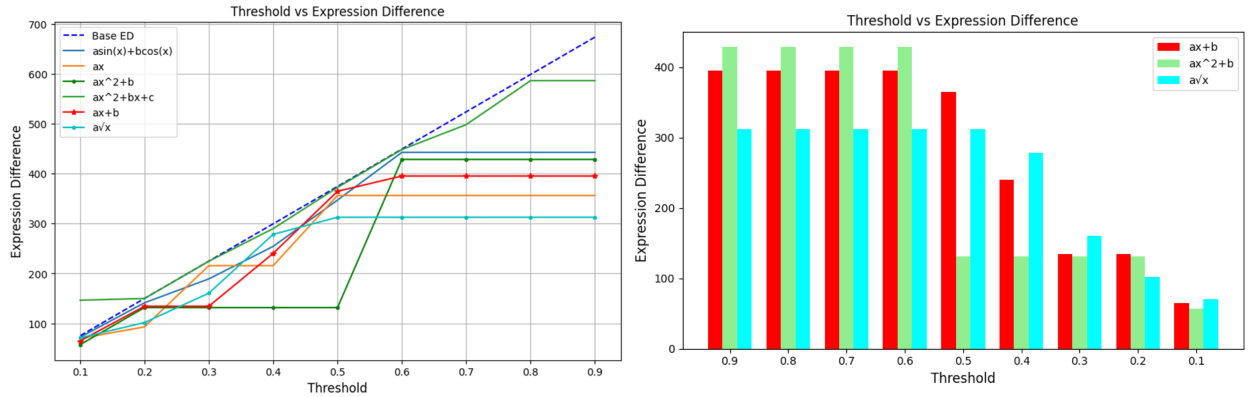
3. **Saccharomyces cerevisiae scRNA-seq data:** Gene expression data from Omnibus GSM3564448 (link) and Omnibus GSM4039308 (link).

4 Results

Here, we present the results of our experiment. We set the threshold as the expression difference between the source and target states and present our results against this threshold.

4.1 Hypoplastic Left Heart Syndrome (HLHS) Results

First, we present the Threshold vs. Expression Difference curve (Figure 2). Here, we observe that as the threshold decreases, the expression difference between the inferred state and the source state also reduces. For the HLHS dataset, the quadratic relationship performs notably well, while the linear relationship, which was the original assumption of GAGER [1], does not perform as effectively.



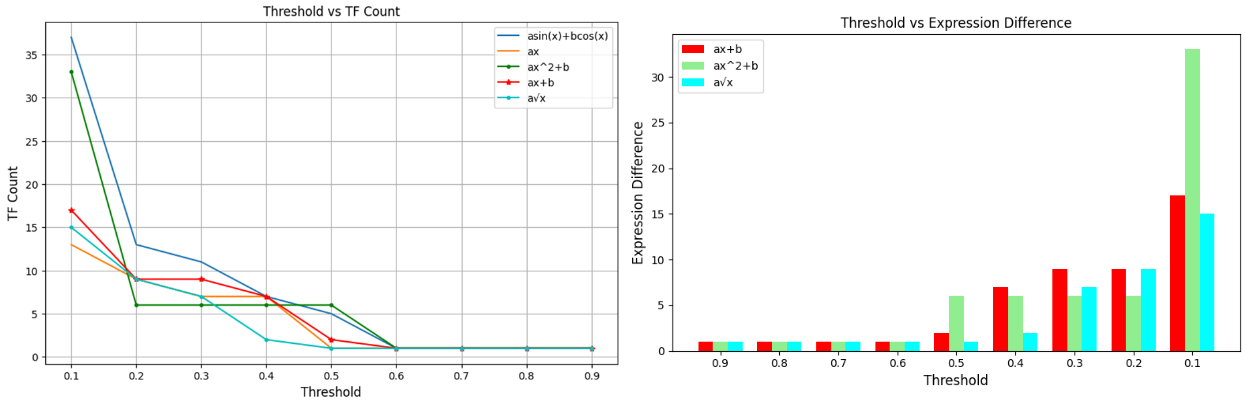
(a)

(b)

Figure 2: HLHS Threshold vs. Expression Difference Plot: The plot shows that the quadratic relationship performs the best.

Next, we present the Threshold vs. TF Count plot (Figure 3). Here, we observe that up to a 20% threshold, the quadratic relationship is effective. However, below 20%, it performs worse than the linear relationship. Interestingly, another nonlinear relationship ($a\sqrt{x}$) always performs better than the linear one. If we look back at the Threshold vs.

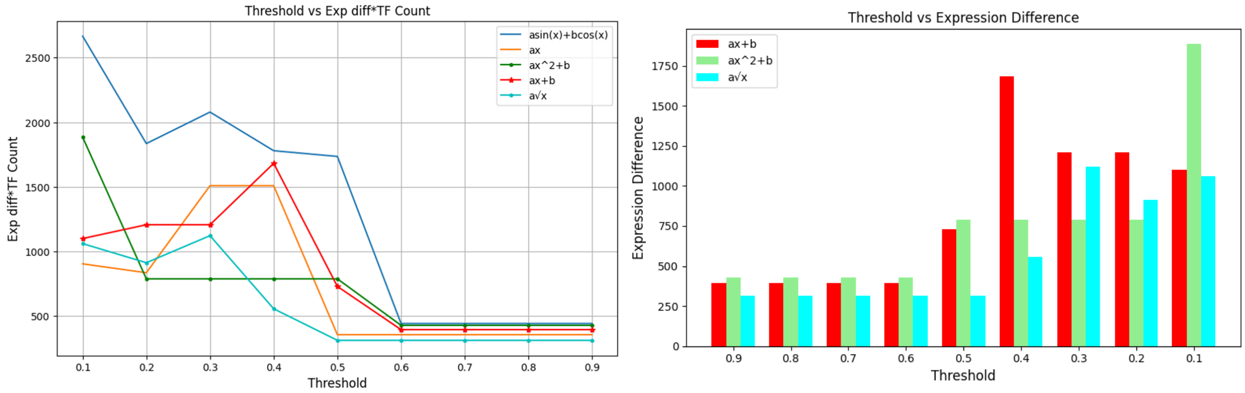
Expression Difference plot (Figure 2), we also notice that the nonlinear relationship ($a\sqrt{x}$) performs better than the linear one.



(a) (b)

Figure 3: HLHS Threshold vs. TF Count: The plot shows that the quadratic relationship performs worse below the 20% threshold, but $a\sqrt{x}$ performs overall better.

As we want to use the minimum number of TFs to restore the expression difference, we also present the Threshold vs. Expression Difference * TF Count plot (Figure 4). Here, we give equal importance to the expression difference and TF count. In this plot, we notice that in three different region, there relationship performs better and none of which correspond to the linear model. From 10% to 20%, the linear relationship with intercept zero performs best; from 20% to 36%, the quadratic relationship is most effective; and above 36%, the $a\sqrt{x}$ relationship performs better. Moreover, we can safely conclude that $a\sqrt{x}$ is the best performing relationship, as its performance does not degrade in the other regions.

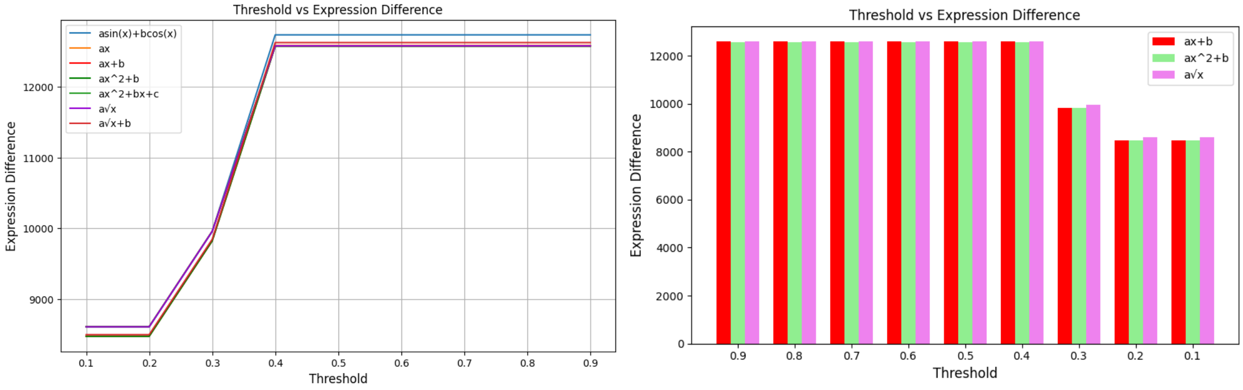


(a) (b)

Figure 4: HLHS Threshold vs. Exp Diff * TF Count: The plot shows that from 10% to 20%, the linear relationship with intercept zero performs best; from 20% to 36%, the quadratic relationship is most effective; and above 36%, the $a\sqrt{x}$ relationship performs better.

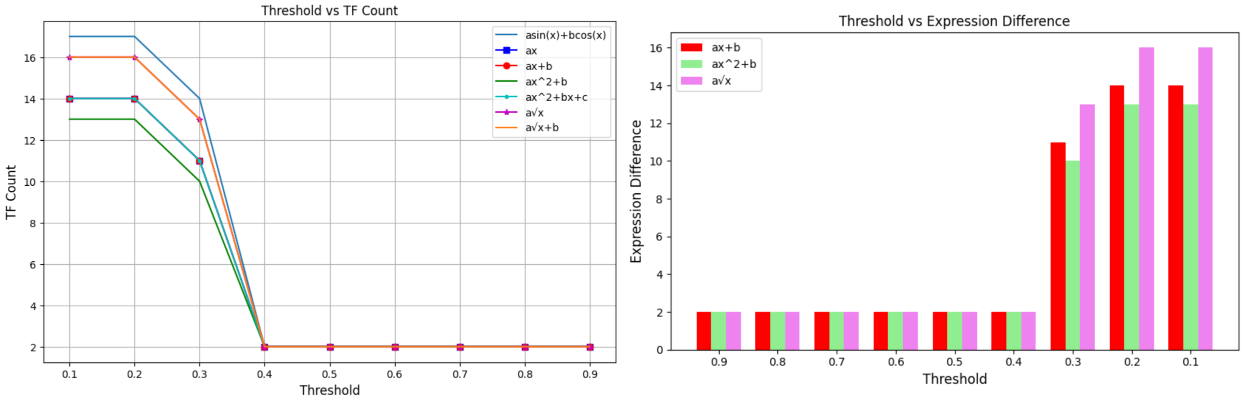
4.2 Progression of Neurodegeneration in Mouse (PNM) Results

We notice somewhat similar result in this dataset as well. Across the overall threshold range, the quadratic relationship performs better (Figure 5, 6, 7). From the Threshold vs. TF count plot (Figure 6) and the Threshold vs. Expression Difference * TF Count plot (Figure 7), we notice that the best-performing $a\sqrt{x}$ relationship in the HLHS dataset does not perform well here.



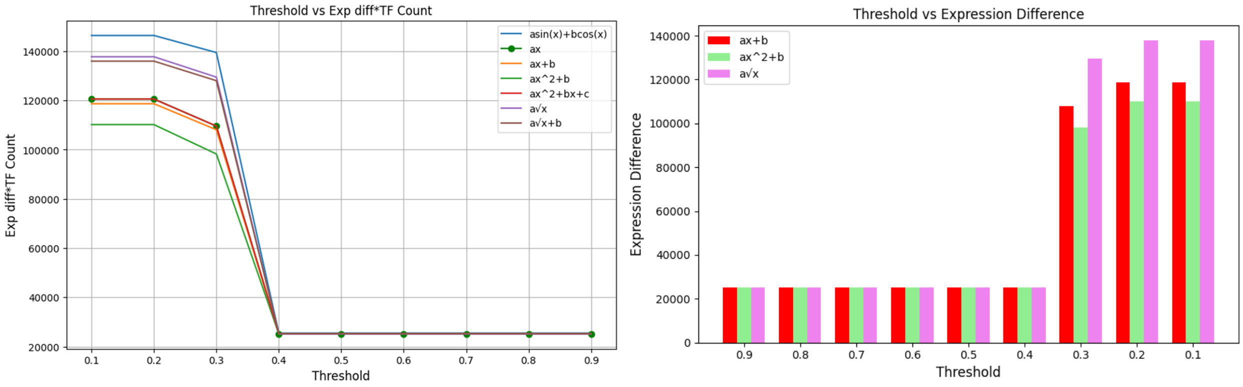
(a) **(b)**

Figure 5: PNM Threshold vs. Expression Difference Plot: The plot shows that all the relationships perform somewhat similarly, but the quadratic and linear relationships perform the best.



(a) **(b)**

Figure 6: PNM Threshold vs. TF Count: The plot shows that the quadratic relationship performs better over all region.

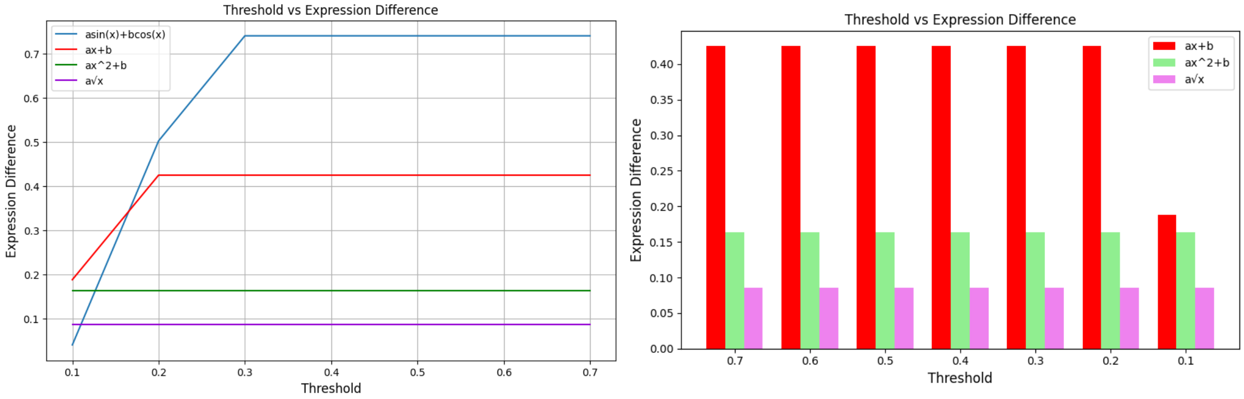


(a) **(b)**

Figure 7: PNM Threshold vs. Exp Diff * TF Count: The plot shows that over all region the quadratic relationship is most effective.

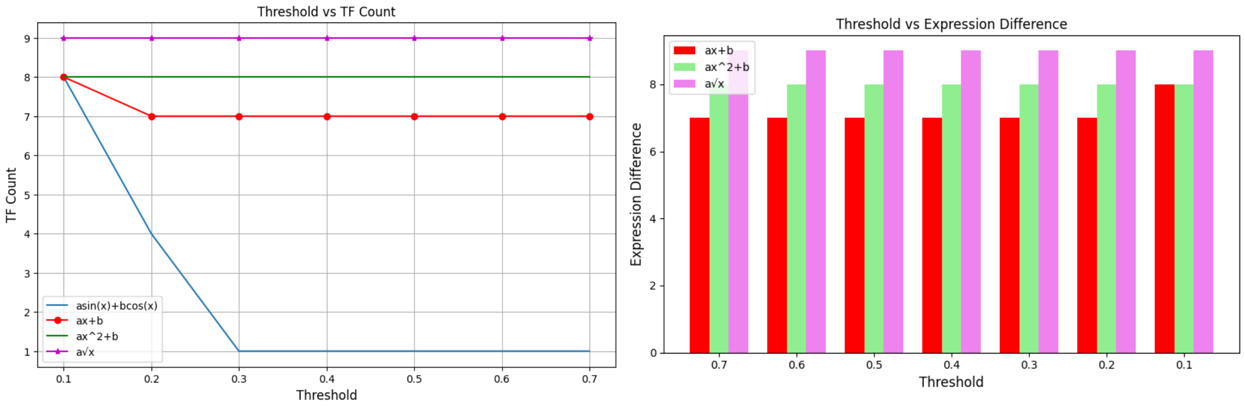
4.3 *Saccharomyces cerevisiae* (Yeast) Results

Here we present the results for *Saccharomyces cerevisiae* dataset (Figure 8, 9, 10). The Threshold vs. TF Count plot (Figure 9) shows that the sinusoidal relationship uses only one TF. However, it is not the best-performing relationship, as it fails to reduce the expression difference (Figure 8).



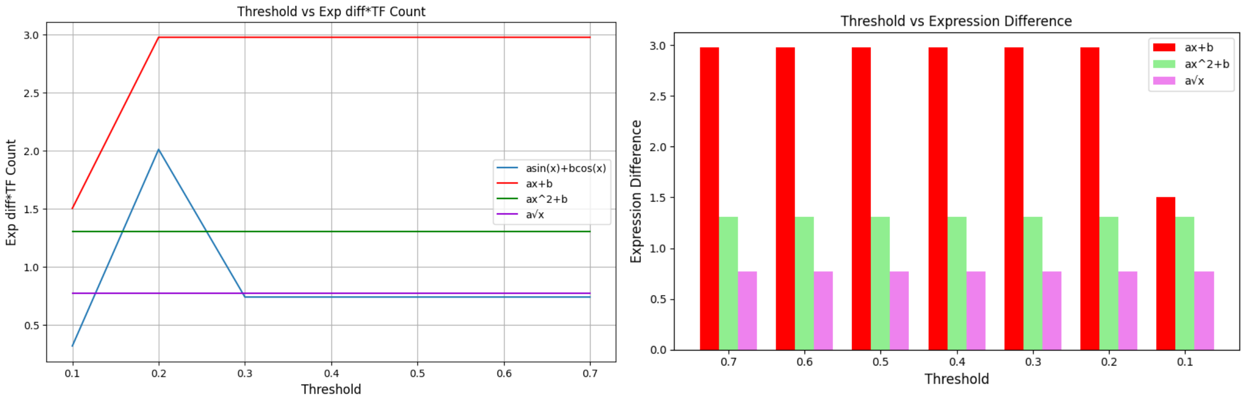
(a) **(b)**

Figure 8: Yeast Threshold vs. Expression Difference Plot: The plot shows $a\sqrt{x}$ relationships perform the best.



(a) **(b)**

Figure 9: Yeast Threshold vs. TF Count



(a) **(b)**

Figure 10: Yeast Threshold vs. Exp Diff * TF Count: $a\sqrt{x}$ is the best performing relationship.

4.4 Summary

The following table summarizes the key results across the datasets.

Dataset	**Threshold**	**Exp Difference**	**TF Count**	**Best Performing Relationship**
HLHS	10%	70.714	15	$a\sqrt{x}$
PN in Mouse	10%	8473.726	13	Quadratic
Yeast	10%	0.085	9	$a\sqrt{x}$

Table 1: Summary of thresholds, expression differences, TF counts, and best-performing relationships across the three datasets.

5 Conclusion

Our project demonstrates that gene expression regulation often involves complex, nonlinear interactions between genes and transcription factors (TFs). Linear models may fail to capture these relationships. By applying nonlinear models across multiple scRNA-seq datasets, we show that nonlinear models can reduce expression differences more effectively while using fewer TFs.

These results indicate that different TF-gene pairs may follow distinct regulatory relationships, and a single model cannot adequately describe all interactions.

A future direction of work is to infer the specific relationship for each TF-gene pair and use these relationships for more accurate gene expression restoration.

References

[1] Chowdhury, M.Z.U.S., Any, S.S., Samee, M.A.H. and Rahman, A. “GAGER: gene regulatory network assisted gene expression restoration.” bioRxiv, pp.2024-11.

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

[3] Davidson, E.H., Erwin, D.H. “Gene regulatory networks and the evolution of animal body plans.” Science 311(5762), 796–800 (2006)

[4] Inoue, M. and Horimoto, K. “Relationship between regulatory pattern of gene expression level and gene function.” PLoSOne. 2017 May 11;12(5):e0177430.

[5] Frank, T.D., Cavadas, M.A., Nguyen, L.K. and Cheong, A., 2016. “nonlinear dynamics in transcriptional regulation: biological logic gates.” In Nonlinear Dynamics in Biological Systems(pp. 43-62). Cham: Springer International Publishing.

[6] Steinacher, A., Bates, D.G., Akman, O.E. and Soyer, O.S., 2016. “Nonlinear dynamics in gene regulation promote robustness and evolvability of gene expression levels.” PloSone, 11(4), p.e0153295.

[7] Gaiteri, C., Ding, Y., French, B., Tseng, G.C., Sibille, E. “Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders.” Genes, brain and behavior 13(1), 13–24 (2014)

[8] Zickenrott, S., Angarica, V., Upadhyaya, B., Del Sol, A. “Prediction of disease–gene–drug relationships following a differential network analysis.” Cell death & disease 7(1), 2040–2040 (2016)

[9] Aibar, S., Gonza´lez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. “Scenic: single-cell regulatory network inference and clustering.” Nature methods 14(11), 1083–1086 (2017)

[10] Love, M.I., Huber, W., Anders, S. “Moderated estimation of fold change and dispersion for rna-seq data with deseq2.” bioRxiv (2014) https://doi.org/10.1101/002832, https://www.biorxiv.org/content/early/2014/11/17/002832.full.pdf

[11] Kamimoto, K., Hoffmann, C.M., Morris, S.A. “Celloracle: Dissecting cell identity via network inference and in silico gene perturbation.” bioRxiv, 2020–02 (2020)