

---

# Learning to Relative Expression under Batch Effects and Stochastic Noise in Spatial Transcriptomics - Supplementary material -

---

Anonymous Author(s)

Affiliation

Address

email

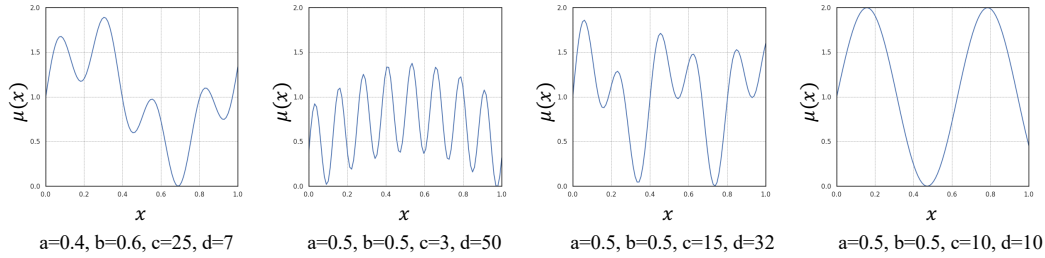


Figure 1: Visualization of  $\mu(x)$ .

## A Details of experiments on synthetic dataset

Figure 1 shows four types of mean functions for our synthetic data. A nonlinear function was chosen to generate waveforms characterized by varying frequencies and slope gradients. This property allows the function to model complex, non-uniform signal behavior, which is relevant in representing heterogeneous patterns observed in gene expression data.

Figure 2 shows the performance of each loss function under various parameters in the synthetic dataset. We changed dispersion parameter  $r$ , scale  $\alpha$ , bias  $\beta$ , scale for tissue 2  $\alpha$ , bias for tissue 2  $\beta$ . Overall, our loss function outperforms the comparisons on each condition. The proposed loss function demonstrates robustness under low-scale conditions. Furthermore, its effectiveness improves as the variability in intensity scales across patients increases.

## B Computer resources

We used the Cloud Environment for the experiment on synthetic data, and an internal desktop computer for Experiment 2.

Experiment 1 (Cloud environment)

- CPU: 16 assigned physical CPU cores
- GPU: None
- Memory: 320 GB

Experiment 2 (Internal desktop environment)

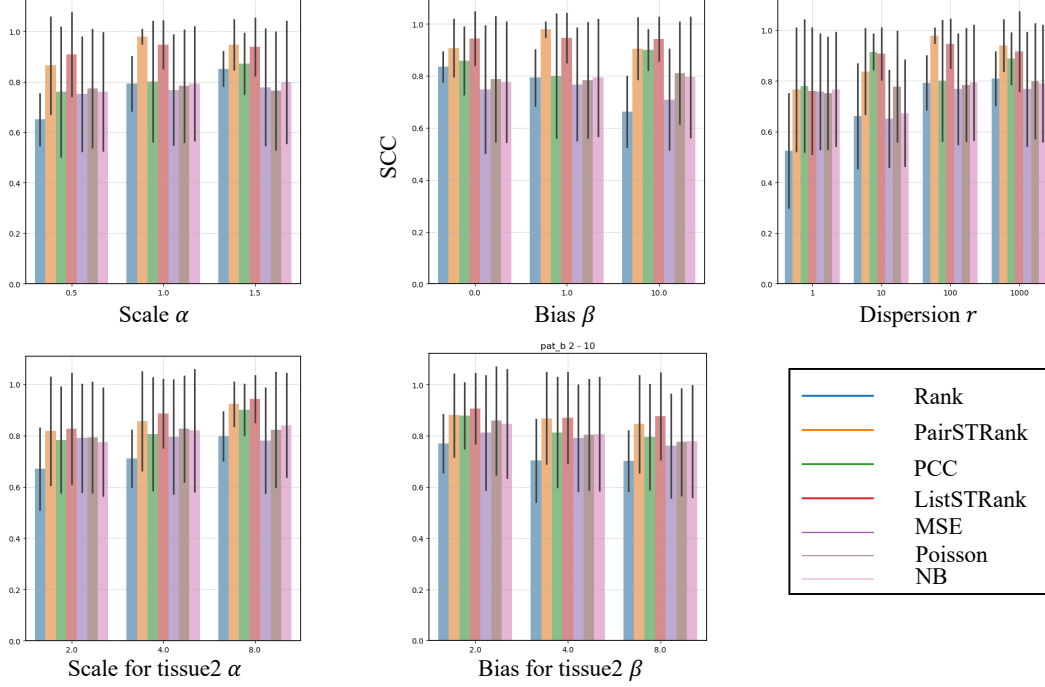


Figure 2: Visualization of  $\mu(x)$ .

- CPU: 12th Gen Intel(R) Core(TM) i9-12900KS, Physical Cores: 16
- GPU: NVIDIA RTX A6000
- Memory: 128 GB

## C Licenses for existing assets

We implemented our method with Pytorch [4] with modified BSD LICENSE, PytorchLightning [1] with Apache-2.0 LICENSE. For the feature extraction from whole slide image, we modified CLAM implementation [3]. For the experiments, we used Hest 1k [2] with CC BY-NC-SA 4.0.

## D Experiments by increasing the number of target genes

Our STRank introduces a probabilistic model, and it is expected to be effective for low signal and sparse conditions. To assess the effectiveness of our probabilistic loss function under this condition, we evaluate its performance as the number of target genes is varied from 50 to 250 by using the Xenium modality of Breast cancer of Hest 1k. We selected the top 50 or 250 genes based on their high variability.

Table 1 summarizes performance comparisons as the number of target genes is varied. Since the gene observation becomes sparse and low signal along with variability, the 250 condition contains many more sparse and low signal genes than the 50 condition. Therefore, we expect the effect of our loss function will be significant on 250 genes. Contrary to our expectations, increasing the number of genes from 50 to 250 did not result in a significant difference. Although our method, which integrates a probabilistic modeling framework, is generally expected to outperform the PCC, the performance advantage diminished with larger gene sets.

This suggests that the reduced performance gap may be attributed to two factors: (1) the test data itself contains stochastic noise, and this effect prevents accurate evaluation, and (2) as the number of highly variable genes increases, noise becomes a dominant factor, thereby diminishing the discriminative signal. To address these issues, it is necessary to consider an evaluation framework; however, this is beyond the scope of the current study.

Table 1: Performance versus number of target genes.

Loss	50	250
PCC	0.498	<u>0.460</u>
PairSTrank	<b>0.520</b>	<b>0.463</b>
ListSTrank	<u>0.516</u>	0.459

## References

- [1] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.
- [2] Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Neurips*, 37:53798–53833, 2024.
- [3] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.
- [4] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.