
Learning Relative Gene Expression Trends from Pathology Images in Spatial Transcriptomics

Kazuya Nishimura¹ Haruka Hirose¹ Ryoma Bise² Kaito Shiku² Yasuhiro Kojima¹

¹ Laboratory of Computational Life Science, National Cancer Center Japan

² Department of Advanced Information Technology, Kyushu University, Japan
kanishi4@ncc.go.jp

Abstract

Gene expression estimation from pathology images has the potential to reduce the RNA sequencing cost. Point-wise loss functions have been widely used to minimize the discrepancy between predicted and absolute gene expression values. However, due to the complexity of the sequencing techniques and intrinsic variability across cells, the observed gene expression contains stochastic noise and batch effects, and estimating the absolute expression values accurately remains a significant challenge. To mitigate this, we propose a novel objective of learning relative expression patterns rather than absolute levels. We assume that the relative expression levels of genes exhibit consistent patterns across independent experiments, even when absolute expression values are affected by batch effects and stochastic noise in tissue samples. Based on the assumption, we model the relation and propose a novel loss function called STRank that is robust to noise and batch effects. Experiments using synthetic datasets and real datasets demonstrate the effectiveness of the proposed method. The code is available at <https://github.com/naivete5656/STRank>.

1 Introduction

With the development of spatial transcriptomic techniques (ST), the comprehensive gene expression profile can be captured on a small spot with a spatial location corresponding to the pathology image [16]. Due to the high cost of acquiring spatial transcriptomics (ST) data, there is growing interest in using computer vision techniques to estimate gene expression from pathology images as a more affordable way [19, 30, 4, 7].

One of the main difficulties in estimating gene expression from pathology images is the batch effects and stochastic fluctuations in observed data. As shown in Figure 1 (a), differences in reagent batches, equipment, and other technical factors in the measurement process (i.e., batch effects) can cause variations in data scaling across tissues [11, 18]. Additionally, due to cellular heterogeneity and temporal dynamics, the observed gene expression level stochastically fluctuates even though the appearance of the pathology image is the same as shown in Figure 1 (b).

Although mean squared error (MSE) is commonly used in the previous estimation methods [7, 19, 30, 4], it is hard to capture variations from the data containing batch effects and stochastic noise. MSE loss focuses on predicting the absolute values of gene expression without correcting for batch effects. Consequently, models trained with MSE loss may inadvertently learn patient-specific biases rather than biologically relevant signals. Additionally, since MSE loss does not model stochastic noise explicitly, it can not account for the significance of biological signals from expression count data.

In this paper, we aim to estimate the relative expression relation instead of directly estimating absolute values of gene expression. The key hypothesis of this paper is that relative gene expression trends

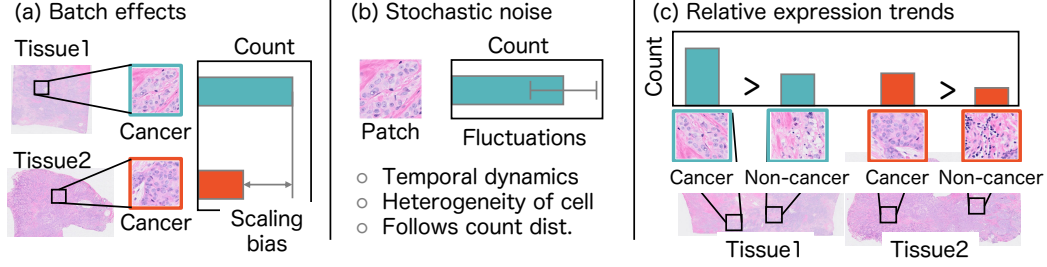


Figure 1: (a) Illustration of scaling bias due to batch effects, (b) stochastic noise, and (c) our hypothesis: learning relative expression trends. Even in the presence of batch effects and stochastic noise, the relative expression trends between patches are preserved.

between image patches are preserved, even when batch effects or stochastic noise are contained in count values. For instance, as illustrated in Figure 1 (c), if we extract cancerous and non-cancerous patches from several tissues, the expression of the cancer cell-specific gene is expected to exhibit higher expression in cancerous patches for all the tissues. Although the absolute expression values and their scales may vary across tissues, we posit that the relative relation of expression between patches remains consistent. In addition, the relative expression relation has been widely used for downstream analysis, such as differential expression analysis, which detects the relative expression difference between clusters. Therefore, capturing relative expression differences between patches within each tissue is more reasonable than directly estimating the absolute value of gene expression.

Learning to rank (i.e., ranking loss) is one of the solutions to learn the relationships between samples [3, 22, 9, 10, 27]. This ranking loss learns which of a given pair of samples has a higher score, and the pairwise learning approach can mitigate batch effects. However, because signal-to-noise ratios of lowly expressed genes tend to be lower than those of highly expressed genes, stochastic noise can alter the relative ranking between samples when gene expression levels are low. Therefore, it is essential to model the relative relationships in a manner that reflects the probabilistic characteristics of gene expression data.

To address the challenge of learning relative relationships in gene expression data affected by stochastic noise, we propose a novel loss function that models gene expression as a discrete probabilistic distribution conditioned on pairwise or listwise input. Specifically, we assume that the expression counts of paired patches are the consequence of the counting process given relative frequencies among the patches; we assume a binomial distribution for pairwise and a multinomial distribution for listwise scenarios. This formulation enables the model to capture relative relationships in a manner that is consistent with the probabilistic nature of observed count data.

To confirm the characteristics of our proposed loss function, we compared a previous loss function with ours using a synthetic dataset. The experiments demonstrated the effectiveness of batch effects and stochastic noise in the low signal situation. Moreover, experiments using real datasets show the generability of our loss function.

Our contributions are highlighted as follows:

- We redefine the task setup of gene expression estimation from pathology images as a rank score estimation setting. The setting is more practical and realistic for the downstream tasks of using gene expression.
- We propose a noise-robust loss function designed to handle batch effects and probabilistic noise, which dynamically adjusts its weighting based on the relative magnitude of expression for spatial transcriptomic data (STRank). This allows the model to learn more effectively, even when the expressions are sparse.
- We demonstrate the effectiveness of our loss function for batch effect and stochastic noise using synthetic datasets. In addition, we validated the robustness of the proposed method in expression estimation using real-world data, confirming its effectiveness under practical conditions.

2 Setup and Notation

Let $\mathcal{X}^{(n)} = \{x^{n,i}\}_{i=1}^{N^p}$ denote a set of image patches in n -th tissue ($n = 1, \dots, N^{(\text{tissue})}$) and $\mathcal{E}^{(n)} = \{e^{n,i}\}_{i=1}^{N^p}$ be a set of their corresponding gene expression levels at each patch, where N^p is the number of patches. The e^i is an N^g -dimensional vector and indicates gene expression level for each gene, where N^g is the number of genes. Unlike previous expression estimation setups that directly estimate expression value $e^{n,i}$ from $x^{n,i}$ without considering tissue n , we aim to learn a function $f : x^{n,i} \rightarrow r^{n,i}$ that estimates the rank score [3] for each gene g , which reflects the relative relation of gene expression between the given patches from the same tissue. The $r_g^{n,i}$ is the scale-invariant, it reflects relative relation among expression of tissue n , where if $e_g^{n,i} > e_g^{n,j}$, then the rank score should be $r_g^{n,i} > r_g^{n,j}$. Since raw gene expression values have some scaling biases introduced by the complexity of the observation technique and experimental conditions, our setup is more intuitive and practical than directly estimating the raw expression value. For readability, we omit the subscript n in, $x^{n,i}$ and $e^{n,i}$. These values are referred to as x^i and e^i , which are samples in tissue n .

To motivate this approach, we first summarise the two losses most widely adopted in prior work — MSE (pointwise) and Rank (pairwise) — and clarify why they remain vulnerable to either batch effects or stochastic noise.

A mean squared error loss between raw expression value e^i and the estimated expression has been widely used for gene expression estimation from pathology image [7, 19, 30]. The mean squared error (MSE) loss for a sample x^i is defined as:

$$L_{\text{MSE}}(e^i, \hat{r}^i) = \frac{1}{N^g} \|e^i - \hat{r}^i\|^2, \quad (1)$$

where $\|\cdot\|$ is the L2 norm, and $\hat{r}^i$ is the output of a function f that estimates the expression value, such as a neural network, $\hat{r}^i = f(x^i)$.

The loss function is calculated based on one patch (*i.e.*, pointwise input). If there are batch effects between $\mathcal{E}^{(n)}$ and $\mathcal{E}^{(m)}$, the loss function can be influenced by scaling bias due to batch effects because the loss does not consider relations between samples obtained from different tissues. When patient data is imbalanced, the model may disproportionately rely on data from certain individuals. This can cause the model to learn spurious patterns, resulting in experimental bias.

To deal with the batch effects in the target data, the pairwise loss, such as Rank loss [3] or listwise loss, such as PCC loss [25, 2] that learn relative relations from multiple inputs, is one of the countermeasures. Ranking loss is designed to capture the relative ordering between pairwise samples. Given pairwise samples from the same tissue n , x^i and x^j , such that their gene expression satisfies $e^i > e^j$, the loss is computed for each pairwise sample in the batch as follows:

$$\mathcal{L}_{\text{Rank}}(\hat{r}^i, \hat{r}^j) = \max(0, \hat{r}_g^i - \hat{r}_g^j + \varepsilon), \quad (2)$$

where ε is a margin value.

While we could learn pairwise relations in tissue by introducing the Rank loss, since the Rank loss does not consider stochastic fluctuation, it is difficult to capture the signal in low signal conditions. The gene expression profile captured by spatial transcriptomics is very sparse and noisy. Therefore, to capture the signal from such a sparse dataset, we should consider the probabilistic model for the loss function.

3 Spatial Transcriptomics Ranking Loss

The motivation of our Spatial Transcriptomics Ranking Loss (STRank) is to learn the relative relationships of gene expression by considering the stochastic noise effect using the distribution of count data by modeling gene expression counts at multiple spots. Similar to the learning to rank setting, we consider two setups: pairwise and listwise loss functions.

3.1 Pairwise STRank Loss

Let us consider the pairwise situation similar to learning to rank [3]. Given a pair of patches i, j obtained from a single tissue, we train the model f to predict the rank score r^i , which reflects

the relation of gene expression e^i, e^j between the samples from patch images x^i, x^j . In contrast to the conventional rank loss function [3] that focuses only on learning ordinal relationships (i.e., which sample is larger), our proposed loss function captures relative differences by incorporating the magnitude of gene expression level.

We assume that the expression count e^i on the spot i , given the pairwise patches x^i, x^j and total expression $t^{i,j}$, follows a Binomial distribution.

$$\Pr(e^i|x^i, x^j, t^{i,j}) = \prod_{g=1}^{N^g} \Pr(e_g^i|x^i, x^j, t_g^{i,j}), \Pr(e_g^i|x^i, x^j, t_g^{i,j}) = \text{Binomial}(t_g^{i,j}, p_g^i), \quad (3)$$

where N^g is the number of genes, p_g^i is the frequency parameter of Binomial distribution, which quantify how frequently the gene g is observed at spot i given the expression count is derived from either spot i or j , and $t_g^{i,j}$ is the total expression level of gene g across these two spots: $t_g^{i,j} = e_g^i + e_g^j$. This modeling approach accounts for the unique statistical characteristics of count data. It enables adaptive weighting of inter-sample relationships based on the observed count levels, thereby improving the ability to learn from count data distributions which has stochastic fluctuations.

Given pairwise patches, x^i and x^j , the model f output scores, $\hat{r}^i$ and $\hat{r}^j$, where $\hat{r}^i, \hat{r}^j \in \mathbb{R}^{N^g}$. A softmax function is then applied between $\hat{r}^i$ and $\hat{r}^j$.

$$\hat{p}_g^i = \frac{\exp(\hat{r}_g^i)}{\exp(\hat{r}_g^i) + \exp(\hat{r}_g^j)}, \quad \hat{p}_g^j = \frac{\exp(\hat{r}_g^j)}{\exp(\hat{r}_g^i) + \exp(\hat{r}_g^j)}, \quad \hat{p}_g^j = 1 - \hat{p}_g^i. \quad (4)$$

Our loss function models the predicted probabilities $\hat{p}_g^i$ and $\hat{p}_g^j$ as parameters of a binomial distribution, and the model is trained by minimizing the negative log-likelihood of the binomial distribution, thereby aligning the predicted distributions with the observed count-based outcomes. The negative log-likelihood can be decomposed into the following form:

$$-\log \Pr(e^i|x^i, x^j, t^{i,j}) = -\log \left(\prod_{g=1}^{N^g} \binom{t_g^{i,j}}{e_g^i} p_g^{i e_g^i} p_g^{j e_g^j} \right) \quad (5)$$

$$= -\sum_{g=1}^{N^g} \left(e_g^i \log p_g^i + e_g^j \log p_g^j + \log \binom{t_g^{i,j}}{e_g^i} \right). \quad (6)$$

Since $\log \binom{t_g^{i,j}}{e_g^i}$ is constant value, our final loss function $L_{\text{STRank}}^{\text{pair}}$ is as follows:

$$L_{\text{STRank}}^{\text{pair}}(x^i, x^j, e^i, e^j) = -\sum_{g=1}^{N^g} (e_g^i \log \hat{p}_g^i + e_g^j \log \hat{p}_g^j). \quad (7)$$

To construct sample pairs, we randomly select a sample from within the same tissue for each reference sample. Patch pairs are generated per tissue using grouped permutation, and the total loss for a mini-batch M , randomly sampled reference without considering patients, is defined to integrate relative signals across tissues as follows:

$$L_{\text{STRank}}^{\text{pair}}(M) = \frac{1}{N^b} \sum_{s=1}^{N^b} L_{\text{STRank}}^{\text{pair}}(x^i, x^{\pi(i)}, e^i, e^{\pi(i)}), \quad (8)$$

where π denotes a permutation index obtained by randomly shuffling the sequential sample indices within each tissue n and $x^{\pi(i)}$ corresponds to a randomly selected sample from the same tissue as x^i . After training, the $\hat{r}$ serves as a rank score indicating the relative expression levels across individual spots.

Relation with Ranking Loss Function. Our proposed pairwise loss function can be interpreted as a relaxed variant of a traditional ranking loss, enabling flexible optimization by considering count value while preserving the core objective of learning relative sample orderings. The previous ranking loss

functions focus solely on the relative ordering between pairs of samples, determining whether one is larger than the other. This implicitly assumes that the difference in rank scores is sufficiently large to make the ordering unambiguous. Under this assumption, $\hat{r}^i$ is sufficiently larger than $\hat{r}^j$, $\hat{r}^i \gg \hat{r}^j$ and the pairwise probabilities are treated as follows: $p_g^i = \frac{\exp(\hat{r}_g^i)}{\exp(\hat{r}_g^i) + \exp(\hat{r}_g^j)} \approx 1$, $p_g^j = \frac{\exp(\hat{r}_g^j)}{\exp(\hat{r}_g^i) + \exp(\hat{r}_g^j)} \approx \frac{\exp(\hat{r}_g^j)}{\exp(\hat{r}_g^i)}$. Then, our loss function can be transformed as follows:

$$L_{\text{STRank}}^{\text{Pair}}(x^i, x^j, e^i, e^j) = - \sum_{g=1}^{N^g} \left(e_g^i \log \frac{\exp(\hat{r}_g^j)}{\exp(\hat{r}_g^i)} \right) = - \sum_{g=1}^{N^g} (e_g^i (\hat{r}_g^j - \hat{r}_g^i)) \propto \hat{r}_g^j - \hat{r}_g^i. \quad (9)$$

Introducing a margin term and a max operation to the difference in rank scores recovers the form of conventional ranking loss functions, such as the hinge-based pairwise ranking loss.

3.2 Listwise STRank Loss

Similar to the pairwise approach, listwise estimation over multiple N^k samples can be formulated by modeling the expression level associated with each sample. This allows the model to handle group-wise comparisons within a unified probabilistic framework.

We assume that the relationship between the list of patch images, $\mathbf{X}^{(n)} = [x^1, \dots, x^{N^k}]$, extracted from the same tissue and their associated gene expression values, $\mathbf{E}^{(n)} = [e^1, \dots, e^{N^k}]$, follows a multinomial distribution. This probabilistic formulation enables modeling the joint contribution of individual patches to the overall expression profile in a listwise manner.

$$\Pr(\mathbf{E}^{(n)} | \mathbf{X}^{(n)}, \mathbf{T}^{(n)}) = \prod_{g=1}^{N^g} \Pr(\mathbf{E}_g^{(n)} | \mathbf{X}^{(n)}, T_g^{(n)}), \Pr(\mathbf{E}_g^{(n)} | \mathbf{X}^{(n)}, T_g^{(n)}) = \text{Multinomial}(T_g^{(n)}, p_g^i), \quad (10)$$

where $T_g^{(n)} = \sum_{i=1}^{N^k} e_g^i$ is the total number of gene expression count, and $\mathbf{E}_g^{(n)} = [e_g^1, \dots, e_g^{N^k}]$. The probabilities for the multinomial distribution p_g^i are obtained using the softmax operation, similar to the pairwise case: $p_g^i = \frac{\exp(\hat{r}_g^i)}{\sum_{j=1}^{N^k} \exp(\hat{r}_g^j)}$.

The negative log likelihood of the multinomial distribution is transformed

$$- \log \Pr(\mathbf{E}^{(n)} | \mathbf{X}^{(n)}, T) = - \log \left(\prod_{g=1}^{N^g} \Pr(\mathbf{E}_g^{(n)} | \mathbf{X}^{(n)}, T_g) \right) \quad (11)$$

$$= - \sum_g \sum_i \left(e_g^i \log p_g^i + \log \frac{T_g!}{e_g^1! e_g^2! \dots e_g^{N^k}!} \right). \quad (12)$$

Since the second term is independent of the model parameters, it can be omitted during optimization. The resulting listwise rank loss for spatial transcriptomics, referred to as ListWiseSTRank, is defined as follows:

$$L_{\text{STRank}}^{\text{List}}(\mathbf{X}^{(n)}, \mathbf{E}^{(n)}) = - \sum_g \sum_i e_g^i \log p_g^i. \quad (13)$$

Similarly to the pairwise loss, we define the total loss for a mini-batch M , which is randomly sampled without considering patients, as follows:

$$L_{\text{STRank}}^{\text{List}}(M) = \sum_{n=1}^{N^{(\text{tissue})}} L_{\text{STRank}}^{\text{List}}(\mathbf{X}^{(n)}(M), \mathbf{E}^{(n)}(M)) \quad (14)$$

where $\mathbf{X}^{(n)}(M)$ and $\mathbf{E}^{(n)}(M)$ are the lists of x and e derived from tissue n in mini-batch M .

Correction Using Expression for Each Spot. Gene expression levels can vary in detectability across spatial spots, so the total count per spot often normalizes expression data. However, such normalization converts the inherently discrete count data into continuous values, which may compromise loss

functions that rely on count-based statistical properties. To mitigate this, we introduce a correction based on the total expression level l^i at each spot, enabling the model to account for inter-spot variability while preserving the count data structure. $p_g^i = \frac{\exp(\hat{r}_g^i) l^i}{\sum_{j=1}^{N^k} \exp(\hat{r}_g^j) l^j}$, where $l^i = \sum_g e_g^i$.

4 Experiments

We evaluated our methods using two types of datasets: synthetic datasets to confirm the hypothesis and characteristics of the proposed loss function, and real datasets to confirm practicality.

Comparisons. We compared our loss function with five loss functions: 1) Mean Squared Error loss (**MSE**) measures the squared difference between the predicted value and the ground truth on a per-sample basis (widely used on gene expression estimation), 2) Poisson loss (**Poisson**) models the output as a Poisson-distributed count and minimizes the corresponding negative log-likelihood for each sample, 3) Negative Binomial loss (**NB**) extends Poisson loss by incorporating a dispersion parameter to handle overdispersed count data at the individual sample level, 4) Rank loss (**Rank**) operates on pairs of samples and penalizes incorrect relative ordering, encouraging proper ranking, 5) Pearson Correlation Coefficient loss (**PCC**) is a listwise loss that maximizes the linear correlation between predicted and true values across the full batch. To examine the effectiveness of our proposed learning relative expression trends strategies, we compare the loss functions of **PairSTRank** (Section 3.1) and **ListSTRank** (Section 3.2), which are based on pairwise and listwise learning, respectively.

The Spearman Correlation Coefficient (SCC) was used as a metric to evaluate the performance.

4.1 Hypothesis Analysis on Synthetic Dataset.

We simulated 1D synthetic data to evaluate the effects of batch effects and stochastic noise. The reason for using synthetic data is that it is difficult to obtain ground truth from raw gene expression datasets since the observed data already contains bias and noise.

In these experiments, each input variable x^i was defined as a one-dimensional scalar constrained to the interval $[0, 1]$. The corresponding gene expression level e^i was modeled using a negative binomial distribution, consistent with prior work in transcriptomic data analysis [11]. Specifically, e^i is also 1D data and was sampled from the distribution $\text{NB}(\alpha\mu(x^i) + \beta, r)$, where $\mu(x^i)$ is the mean response function of the input, r is the dispersion parameter, α is scaling parameter, β is bias parameter.

The objective of the experiments is to accurately estimate the mean function $\mu(x^i)$ from a given dataset $D = \{\mathcal{X}^{(n)}, \mathcal{E}^{(n)}\}$, $\mathcal{X}^{(n)} = \{x^i\}_{i=1}^{N^n}$, $\mathcal{E}^{(n)} = \{e^i\}_{i=1}^{N^n}$. It corresponds to finding the meaningful signal from observed data. The $\mu(x^i)$ is a nonlinear function: $\mu(x^i) = a \sin(cx^i) + b \sin(dx^i) + a + b$, which is the same with [21]. We prepared four types of functions with different parameters (cf. Supplementary material A). The mean SCC for the four types of functions is calculated.

We assumed a training dataset comprising two patients ($n = 2$), and each patient’s gene expression is affected by experimental batch effects, which stem from sources such as differences in imaging protocols or acquisition equipment. We assumed that gene expression measurements for each patient are affected by distinct batch parameters α (scaling) and β (offset). For simplicity, we set $\alpha = 1$ and $\beta = 0$ for the first patient, $\alpha = 10$ and $\beta = 10$, for the second patient (Other conditions are shown in supplementary material A). For training, 50,000 samples were independently sampled from each patient. The validation and test sets, each consisting of 10,000 samples, are sampled from a uniform distribution over the interval $[0, 1]$ in both setups. We compared a uniform situation, where input samples are sampled from a uniform distribution (Figure 2 (a)), and an imbalanced situation, where the sample of tissue 2 is sampled in the specific section (Figure 2 (b)). Figure 2 shows an example of synthetic data. The color indicates tissue, the dotted line indicates the mean function, and each plot represents an observed value e^i , which is sampled with the negative binomial distribution.

A simple MLP (multi-layer perceptron) with 3 linear layers ($[1 \times 128], [128 \times 128], [128 \times 1]$) with ReLU was used for the model. The epoch was 2000 using AdamW [13] with a learning rate $1e-3$ with mini batch size = 256. For the scheduler, we used CosineAnnealing [12]. Details regarding computational resources and related settings are provided in the supplementary material B.

Table 1 shows the comparative performance of various loss functions used for gene expression estimation. Traditional pointwise losses, including widely used Mean Squared Error (MSE), optimize

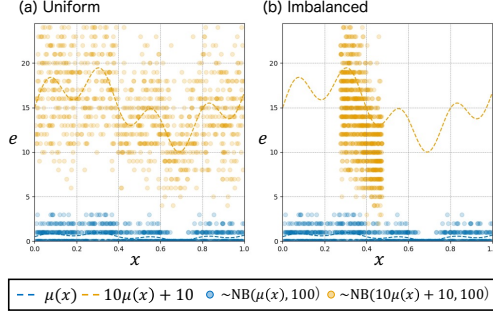


Figure 2: Example of synthetic data for validating the batch effect. Colors indicate patients; the dashed line represents the mean function to be learned; and the dots show observations. (a) Uniform setting: Each patient’s data is drawn from a uniform distribution. (b) Imbalanced setting: Observed data is skewed.

Table 1: Performance comparison across multiple patient conditions. Results are based on synthetic data. Bold faces indicate the best performance in each setting, while underlined values denote the second-best.

		Uniform	Imbalanced
List Pair Point	MSE	0.748	0.583
	Po	0.777	0.603
	NB	0.788	0.601
	Rank	0.835	0.738
	PairSTRank	0.907	0.818
	PCC	0.858	0.560
	ListSTRank	0.945	0.828

prediction accuracy by minimizing the absolute difference in expression levels across individual samples. In contrast, loss functions based on pairwise and listwise learning paradigms, which capture relative expression relationships between genes or samples, demonstrate superior performance overall. These results demonstrate that learning relative expression is effective in situations where batch effects are present. Notably, the proposed methods—PairSTRank and ListSTRank—consistently outperform both conventional pairwise and listwise approaches, indicating their enhanced capacity to model the structured dependencies inherent in gene expression data in both uniform and imbalanced situations. Empirical results show that ListSTRank outperforms its pairwise counterpart. We attribute this improvement to ListSTRank’s ability to capture global expression patterns across entire batches, as opposed to the localized comparisons used in pairwise learning. This suggests that leveraging broader relational context is advantageous under batch-affected conditions.

4.2 Evaluation on Real Datasets

Dataset. To evaluate the effectiveness of our proposed method, we performed experiments using seven datasets from the benchmark of the HEST-1k dataset [8]: IDC, PRAD, PAAD, COAD, READ, ccRCC, and IDC-LymphNode. SKCM and LUAD datasets were excluded from the analysis because they contain only two patients and do not align with the assumptions of our study. Each dataset contains samples from three individual patients. The IDC, PAAD, and COAD datasets were acquired using the Xenium platform, whereas other datasets were obtained via the Visium platform. To avoid train/test patient-level data leakage, we used patient-stratified splits and one patient for validation and testing data, respectively, and the other patients were used for training data. The motivation for the experiment is not to compare models, but to compare loss functions. Therefore, we do not use PCC or regularization, and simply train a regression model with each loss function.

We used 50 genes with highly variable genes. To assess the influence of the loss function, we kept the feature extractor fixed and trained only a single fully connected (fc) layer, as shown in Figure 3. The feature extractor was CONCH [14], which is a vision and language foundation model for pathology. Model optimization employed the AdamW optimizer with a learning rate of $5e - 5$ and a batch size of 256. We trained the model for up to 1000 epochs, with early stopping implemented using a patience threshold of 30 epochs.

Table 2 summarizes the performance comparison across all datasets. Overall, the proposed method outperforms conventional loss functions on average. Although STRank demonstrated superior performance on synthetic data, STRank did not consistently outperform alternatives across all conditions in real datasets. Because real data evaluations are based on observations that inherently include stochastic and measurement noise, it remains essential to assess whether the evaluations reliably reflect true model performance. Even under such conditions, STRank remained relatively stable and was able to demonstrate superior performance on average.

Table 2: Real dataset which is obtained from HEST-1k [8]. Bold faces indicate the best performance in each setting, while underlined values denote the second-best. Ave. is average performance.

	Loss	IDC	PRAD	PAAD	COAD	READ	ccRCC	IDC-L	Ave.
Point	MSE	0.393	0.484	0.307	0.556	0.140	0.093	0.168	0.306
	Po	0.314	<u>0.485</u>	0.336	0.524	0.172	0.091	0.134	0.293
	NB	0.199	0.491	0.119	0.538	<u>0.160</u>	0.075	0.126	0.244
Pair	Rank	<u>0.317</u>	<u>0.317</u>	<u>0.181</u>	<u>0.566</u>	<u>0.047</u>	<u>0.059</u>	<u>0.110</u>	<u>0.228</u>
	PairSTRank	0.494	0.458	0.346	<u>0.613</u>	0.136	0.127	<u>0.228</u>	<u>0.343</u>
List	PCC	<u>0.472</u>	<u>0.459</u>	<u>0.307</u>	0.640	<u>0.105</u>	<u>0.102</u>	<u>0.198</u>	<u>0.326</u>
	ListSTRank	0.510	0.459	<u>0.343</u>	0.597	0.140	<u>0.125</u>	0.238	0.345

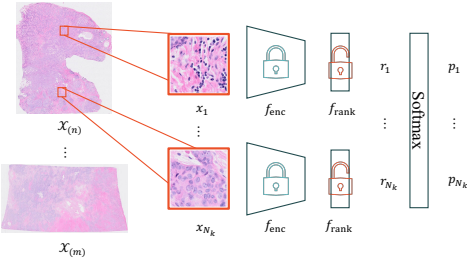


Figure 3: Illustration of our framework for the real dataset. To assess the loss function, we only update the classifier head for this evaluation.

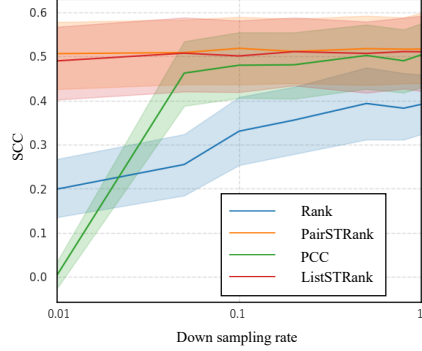


Figure 4: Effectiveness for sparsity. The x-axis is on a log scale.

4.3 Effectiveness for Sparsity

To further assess the robustness of STRank for the sparsity, we conducted performance variability for varying sparsities in a modified real dataset. One way to assess robustness for the sparsity is by assessing the performance on low-expressed genes. However, as discussed in [17], sparsity evaluation becomes challenging in low-expression data due to the lack of known ground-truth signals. To address this, we simulated the sparsity-enhanced expression data for genes with the top 50 highest mean expression levels. For a count of each gene in each cell, we conducted binomial sampling using a specified down-sampling rate to acquire down-sampled count data. We varied levels of expression by downsampling each gene expression count using probabilities $p = 0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 1$.

Figure 4 shows the performance of the pairwise and listwise loss functions on each downsampling rate. Our Pair and List STRank loss outperforms Rank and PCC on each down-sampling rate. The difference becomes significant when the rate is 0.01, effectiveness on highly sparse and weak signals. Since gene expression is inherently sparse, these results suggest that the STRank is well-suited for capturing gene expression signals.

4.4 Effect for Parameter N^k of Our Loss Function

We varied N^k and assessed the effect of N^k , which is the number of samples to calculate our ListSTRank. A larger N^k is expected to be generally preferable, as it facilitates the capture of global trends. However, in the presence of noise or distortion in the global structure, pairwise learning may offer improved performance.

Table 3 shows the performance on each N^k . The results indicate that increasing k beyond 4 leads to improved performance. However, in practice, increasing N^k is not always beneficial due to numerical instability and computational overhead in both situations. For $k \geq 4$, the model demonstrates robustness, with stable performance observed at $k = 8$ and $k = 16$.

Table 3: Performance of STRank for N^k on three conditions.

N^k	Uniform	Imbalanced	B Xenium
2	0.907	0.818	0.447
4	0.938	0.837	0.462
8	0.958	0.839	0.455
16	0.943	0.833	0.458
32	0.938	0.818	0.463
64	0.926	0.827	0.462
128	0.941	0.845	0.457
256	0.945	0.828	0.459

5 Related work

Gene expression estimation from pathological image. Estimating gene expression from pathology images has the potential to reduce sequencing costs and help understand diseases. Deep learning has been introduced in this field, and the deep learning models are trained with patch and gene expression pairs captured by spatial transcriptomics. ST-Net [7] has introduced a transfer learning approach and estimates gene expression by a convolutional neural network pre-trained on ImageNet [5]. To utilize global information of patches, graph convolution neural networks and transformers have been introduced in Hist2gene [19] and Hist2st [30]. To effectively combine local and global information, M2OST [24] and TRIPLEX [4] effectively combined multiple features that are extracted from multiple resolutions. By focusing on the difficulty of directly estimating multiple-dimensional gene expression, exemplar-guided estimation [26, 28, 29], which utilizes retrieved gene expression, has been proposed. BLEEP [26] has trained a model with image and gene expression in a contrastive learning manner and retrieves gene expression based on the image. EGN [29] has refined the retrieved gene expression with a transformer block.

MSE loss has been mainly used as the loss function for these methods. In contrast, we focus on the relative relation among tissues and propose a novel noise-robust loss function with pairwise learning.

Learning to rank. Learning to rank is the field that learns the ranking function from ground-truth rankings [3]. Sculley has adapted the Stochastic Gradient Descent method for learning to rank [22], allowing models to be trained on large datasets. The ranking loss has been integrated with neural networks, which have been widely utilized in applications including Image Quality Assessment [9] and crowd counting [27, 10].

In contrast to these works, which only consider ranking, our loss function considers the size of the count value and adaptively weights depending on the relation of input samples. Since the gene expression is sparse and has a low signal, taking into account the count value helps to learn the relation in the low signal situation.

6 Conclusion

In this paper, we tackled gene expression estimation from pathology images by reconsidering the objective. In contrast to the previous method, which estimates the absolute expression value, we aim to learn the relative gene expression relation. In addition, we propose a novel loss function (STRank) designed to capture the relative gene expression across spatial patches by modeling the relative expression relation. Through comprehensive experiments on both synthetic and real-world datasets, we demonstrated that our method achieves more stable and reliable performance compared to traditional point-wise approaches. These results suggest that exploiting relative gene expression patterns is a promising strategy for enhancing robustness in gene expression prediction.

7 Limitations

We hypothesized that STRank would perform well under sparse conditions, and we evaluated its performance by varying the number of target genes from 50 to all possible genes. As shown in Supplementary D, we could not confirm the effectiveness of STRank in a sparse situation of a real dataset. Furthermore, as shown in Table 2, rigorous evaluation on real data remains an open challenge,

and the fundamental reasons why STRank fails to outperform conventional loss functions in some datasets have yet to be elucidated.

Although STRank shows effectiveness across multiple patients, it converges more slowly than PCC when evaluated on data from a single patient. This behavior is likely due to STRank’s conservative weight updates under low-sample conditions. As a result, PCC may offer better convergence performance on simpler datasets with limited observations.

Another challenge is performance in multi-cohort settings, where the sample is obtained from different experimental conditions (e.g., different hospitals and procedures). As detailed in Supplementary D, our proposed loss functions exhibit limitations when applied to such multi-cohort datasets with a large number of genes. In such cases, batch effects in gene expression are amplified, and pathology images are likewise influenced by experimental batch effects due to variations in imaging conditions. While addressing this issue is beyond the scope of our study, it represents an open question for this field. Future research should aim to develop more robust and generalizable approaches to mitigate these effects.

Acknowledgments and Disclosure of Funding

Funding in direct support of this work: JSPS KAKEN JP24KJ2205, JPMJBS2406, JP23K18509, and JP25K22846, Development (AMED) grant 24ama221609h0001(P-PROMOTE) (to YK), National Cancer Center Research and Development Fund 2024-A-6 (to YK). We used ABCI 3.0 provided by AIST and AIST Solutions.

References

- [1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast tumors reveals novel intercellular relationships. *bioRxiv*, pages 2020–07, 2020.
- [2] Chaoyu Chen, Xin Yang, Ruobing Huang, Xindi Hu, Yankai Huang, Xiduo Lu, Xinrui Zhou, Mingyuan Luo, Yinyu Ye, Xue Shuang, et al. Fine-grained correlation loss for regression. In *MICCAI*, pages 663–672. Springer, 2022.
- [3] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. *Neurips*, 22, 2009.
- [4] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *CVPR*, pages 11591–11600, 2024.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.
- [6] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.
- [7] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.
- [8] Guillaume Jaume, Paul Doucet, Andrew Song, Ming Yang Lu, Cristina Almagro Pérez, Sophia Wagner, Anurag Vaidya, Richard Chen, Drew Williamson, Ahnong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Neurips*, 37:53798–53833, 2024.
- [9] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *ICCV*, pages 1040–1049, 2017.
- [10] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, pages 7661–7669, 2018.

- [11] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- [12] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.
- [13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- [14] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.
- [15] 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.
- [16] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.
- [17] Gabriel Mejia, Daniela Ruiz, Paula Cárdenas, Leonardo Manrique, Daniela Vega, and Pablo Arbeláez. Enhancing gene expression prediction from histology images with spatial transcriptomics completion. In *MICCAI*, pages 91–101. Springer, 2024.
- [18] Kazuya Nishimura, Ryoma Bise, and Yasuhiro Kojima. Towards spatial transcriptomics-guided pathological image recognition with batch-agnostic encoder. In *ISBI*, pages 1–5. IEEE, 2025.
- [19] Minking Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11, 2021.
- [20] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [21] Silvia L Pintea, Yancong Lin, Jouke Dijkstra, and Jan C van Gemert. A step towards understanding why classification helps regression. In *ICCV*, pages 19972–19981, 2023.
- [22] D Sculley. Large scale learning to rank. In *NIPS Workshop on Advances in Ranking*, volume 3, 2009.
- [23] Ryousei Takano, Shinichiro Takizawa, Yusuke Tanimura, Hidemoto Nakada, and Hirotaka Ogawa. Abci 3.0: Evolution of the leading ai infrastructure in japan, 2024.
- [24] Hongyi Wang, Xiuju Du, Jing Liu, Shuyi Ouyang, Yen-Wei Chen, and Lanfen Lin. M2ost: Many-to-one regression for predicting spatial transcriptomics from digital pathology images. *arXiv preprint arXiv:2409.15092*, 2024.
- [25] Ken CL Wong and Mehdi Moradi. 3d segmentation with fully trainable gabor kernels and pearson’s correlation coefficient. In *International Workshop on Machine Learning in Medical Imaging*, pages 53–61. Springer, 2022.
- [26] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Neurips*, 36:70626–70637, 2023.
- [27] Zheng Xiong, Liangyu Chai, Wenxi Liu, Yongtuo Liu, Sucheng Ren, and Shengfeng He. Glance to count: Learning to rank with anchors for weakly-supervised crowd counting. In *WACV*, pages 343–352, 2024.
- [28] Yan Yang, Md Zakir Hossain, Eric Stone, and Shafin Rahman. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition*, 145:109966, 2024.
- [29] Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *WACV*, pages 5039–5048, 2023.

- [30] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *bioRxiv*, 2022.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We clearly described our claims of learning to relative gene expression, and the experiments validated our assumptions.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The manuscript includes a dedicated section discussing the limitations of the study, which are addressed in an appropriate and transparent manner.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our result does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will release the source code upon publication. The details of the experiment conditions were described in the experiment session.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the source code upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We described the details of the implementation detail in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our figure contains error bars, and the variance of the main result is shown in the Supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The details of computer resources are described in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We used publicly available dataset: Hest 1k with CC BY-NC-SA 4.0 LICENSE and

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The work was conducted on fully anonymized data and focuses on methodological development. It does not raise any immediate ethical, legal, or social concerns, and no direct societal impact—positive or negative—is expected at this stage.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not use such types of data and models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All methods employed in this study are appropriately cited. Additional details, including licensing information, are provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We will release the source code upon publication.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: Our experiments used only publicly available datasets.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We used LLMs for only writing, editing, and coding.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Details of Experiments on Synthetic Dataset

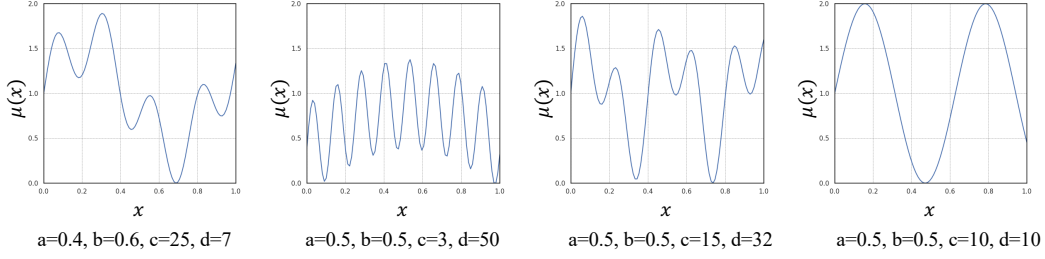


Figure 5: Visualization of $\mu(x)$.

Figure 5 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 the gene expression data.

Figure 6 shows the performance of each loss function under various parameters in the synthetic dataset. We changed scale α , bias β , dispersion parameter r , scale for tissue 2 α , bias for tissue 2 β . Overall, our loss function outperforms all 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.

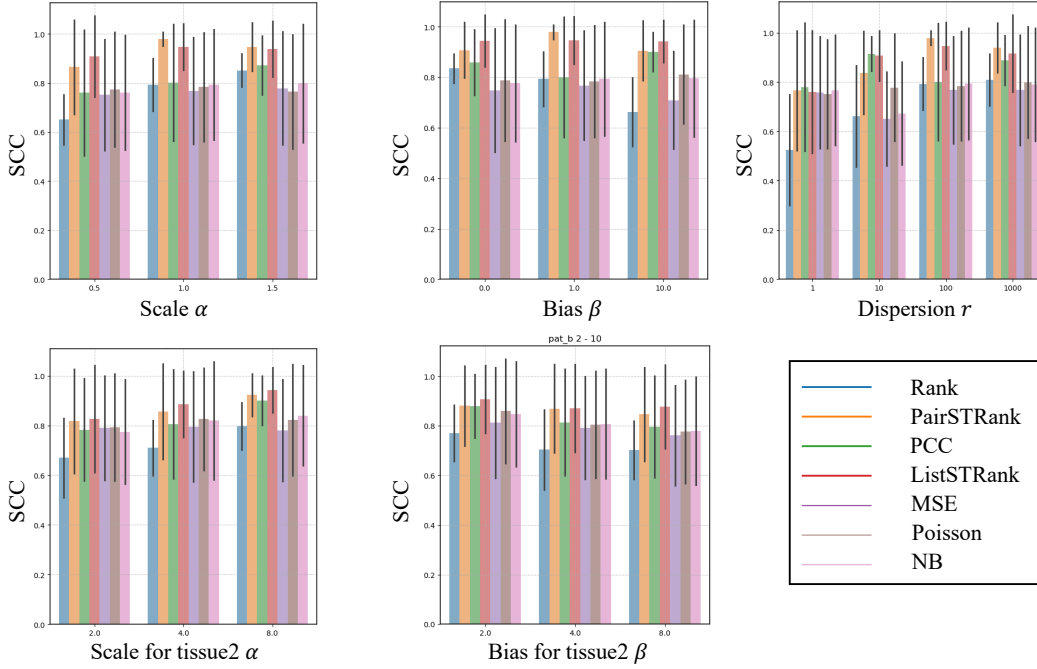


Figure 6: Visualization of $\mu(x)$.

B Computer Resources

We used the Cloud Environment [23] for the experiment on synthetic data, and an internal desktop computer for Experiment 2.

Experiment 1 (Cloud environment)

- CPU: 16 assigned physical CPU cores

Table 4: Performance on HER2ST dataset with 50–9385 gene sets.

Method	50	250	1000	5000	9385
MSE	0.193	0.181	0.172	0.162	0.132
NB	0.020	0.154	0.098	0.083	0.074
Po	0.009	0.150	0.095	0.076	0.069
Rank	0.095	0.052	0.041	0.042	0.018
PairSTRank	0.244	0.194	0.176	0.177	0.173
PCC	0.189	0.173	0.165	0.171	0.152
ListSTRank	0.260	0.175	0.110	0.085	0.087

Table 5: Performance on COAD Visium dataset with 50–2000 gene sets.

Method	50	250	1000	2000
MSE	0.2712	0.2090	0.2125	0.2084
NB	0.3043	0.1682	0.0899	0.0868
Po	0.2572	0.1440	0.0858	0.0829
Rank	0.1118	0.0714	0.0260	0.0599
PairSTRank	0.3456	0.1957	0.1383	0.1319
PCC	0.2615	0.2000	0.2035	0.1951
ListSTRank	0.3398	0.1973	0.1404	0.1375

- GPU: None
- Memory: 320 GB

Experiment 2 (Internal desktop environment)

- 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 [20] with modified BSD LICENSE, PytorchLightning [6] with Apache-2.0 LICENSE. For the feature extraction from whole slide images, we modified the CLAM implementation [15]. We used Hest 1k [8] with CC BY-NC-SA 4.0 for the real datasets.

D Experiments on More Realistic Scenarios

We conducted experiments using the HER2ST datasets [1] and the COAD Visium (multi-cohort) with large gene sets to evaluate the robustness of our loss function on more realistic gene expression estimation scenarios. For HER2ST, we used gene sets of size 50, 250, 1000, 5000, and the full set of 9,385 genes. For COAD, we evaluated the loss functions on gene sets of size 50, 250, 1000, and 2000.

The results are shown in Tables 4 and 5. Our loss function demonstrated robustness on the HER2ST dataset, which is a highly sparse spatial transcriptomics dataset, outperforming other loss functions. This suggests that the proposed loss function is effective for sparse real-world data and for mitigating batch effects within the same cohort. However, for the COAD dataset, our loss function underperformed MSE and PCC when using multi-cohort data with large gene sets (250–2000 genes).

In multi-cohort settings (*e.g.*, COAD setting in Table 5) with many target genes, we assume that the distribution of stochastic noise varies substantially across cohorts; therefore, for low-signal genes where the noise component is dominant, further extensions that account for cohort-specific noise characteristics are needed. To effectively handle multi-cohort data, additional factors beyond batch

Table 6: Results of different methods on Hest2gene

Method	Hest2gene
MSE	0.039
NB	0.005
Po	-0.014
Rank	0.042
PairSTrank	0.052
PCC	0.039
ListSTrank	0.046

effects should be considered, including potential differences in biological signals and domain shifts in image features. This is one of the open problems and is essential for the practical application of gene expression prediction.

E Integration with Previous Method

To evaluate the generalizability of the proposed loss function, we examined its performance when integrated into existing gene expression estimation methods. We evaluated the Spearman correlation coefficient (SCC) for predicting 250 genes using the HER2ST dataset [1]. In contrast to the original study, we adopted a patient-level data split to evaluate generalizability. We then assessed the performance of HisToGene [19] trained with various loss functions. As summarized in the Table 6, the proposed PairSTrank loss achieved the highest performance. In conclusion, our loss functions demonstrated effectiveness on this setup.

F Exploring the Impact of Mini-Batch Sampling Strategies

To investigate the effect of a mini-batch sampling strategy, we compared three settings: the original implementation (Default), a setting where the mini-batch contains samples from only one tissue (Intra-tissue), and a setting where samples are evenly sampled from all tissues (Inter-tissue).

Table 7 shows the results of these experiments on the 250 and 1000 gene sets. Even when we use different mini-batch strategies, the performance of ListSTRank is not improved. PairSTRank loss also does not consider the intra-tissue and inter-tissue, but it still outperforms ListSTRank loss. This suggests that the performance degradation of ListSTRank is not due to the mini-batch setting. Based on these results, we suspect that the performance degradation of ListSTRank may be due to numerical effects.

Table 7: Effect of mini-batch sampling on 250 and 1000. We compare three settings: Default, Intra-tissue (mini-batch contains samples from only one tissue), and Inter-tissue (samples are evenly sampled from all tissues).

Method	250	1000
Default	0.175	0.110
Intra-tissue	0.177	0.105
Inter-tissue	0.173	0.105