Learning Relative Gene Expression Trends from Pathology Images in Spatial Transcriptomics

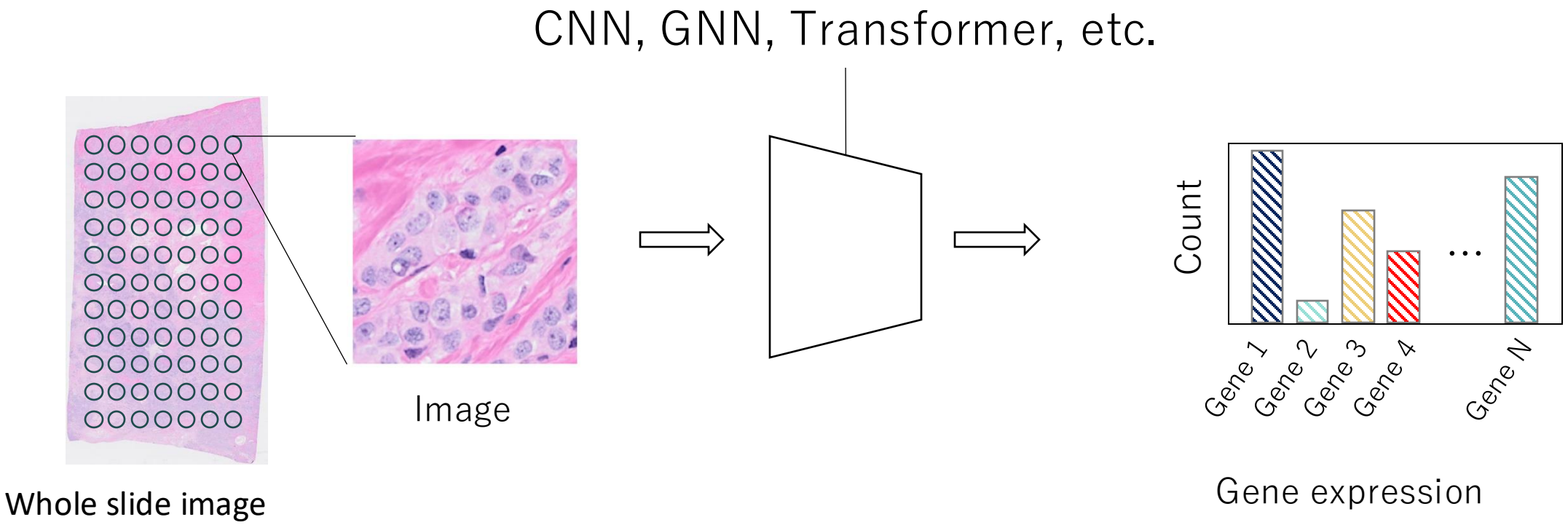
Kazuya Nishimura[1,2], Haruka Hirose[2], Ryoma Bise[3], Kaito Shiku[3], Yasuhiro Kojima[2]

1 Osaka University,

2 National Cancer Center Japan,

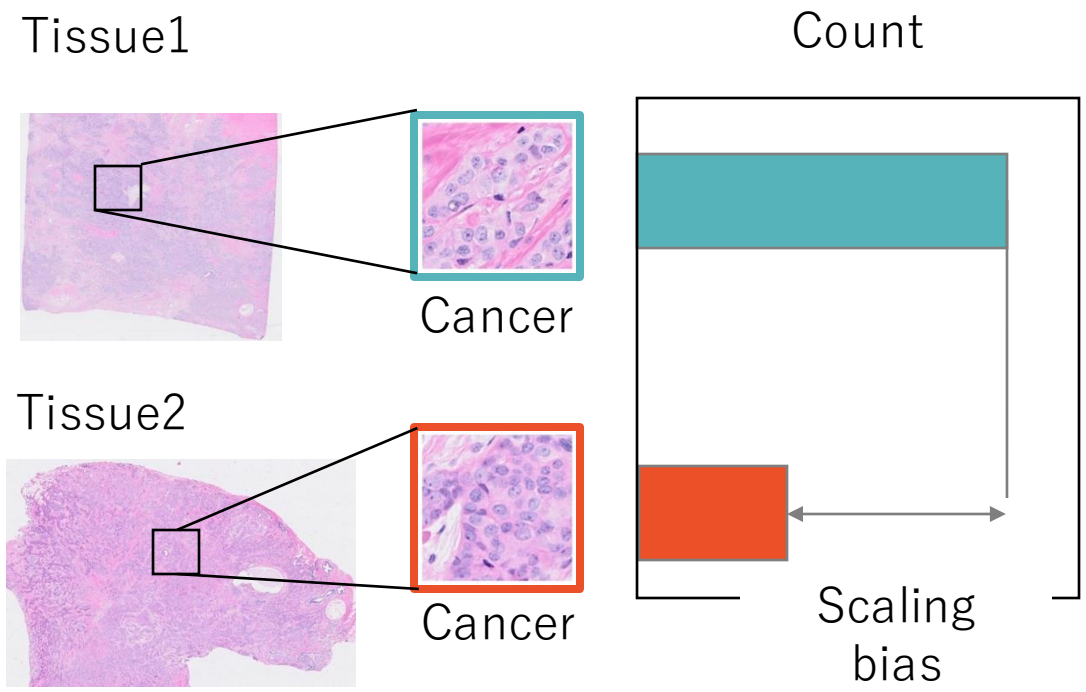
3 Kyushu University

Patch-level gene expression estimation



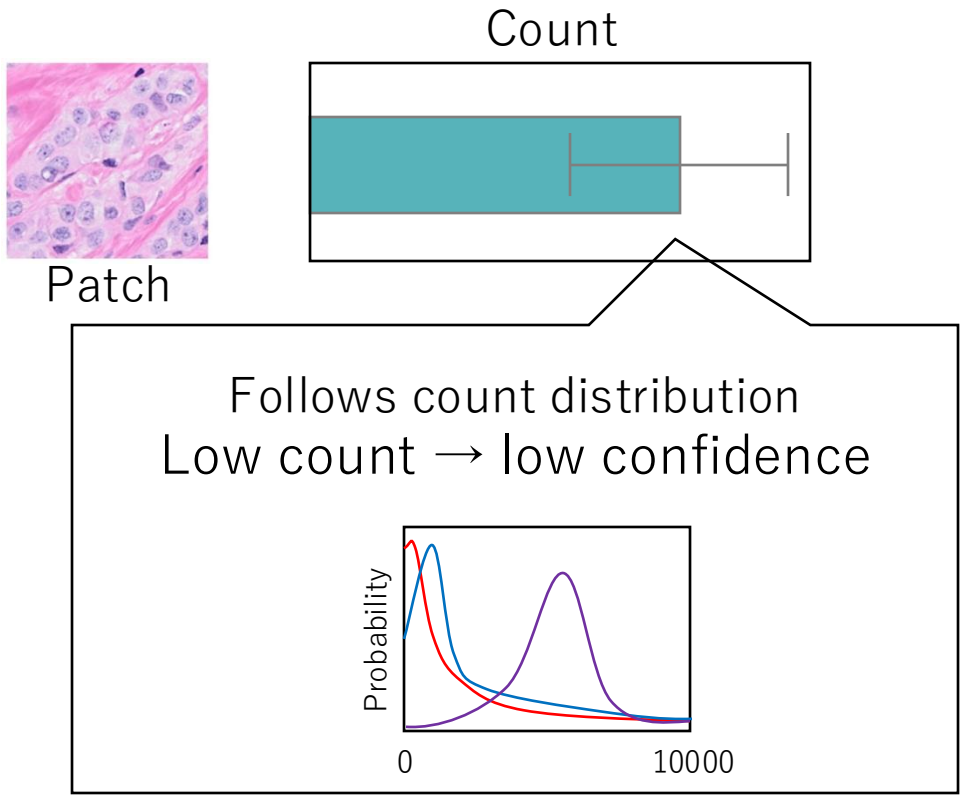
Noise and bias in observed expression

Scaling bias (Batch effect)

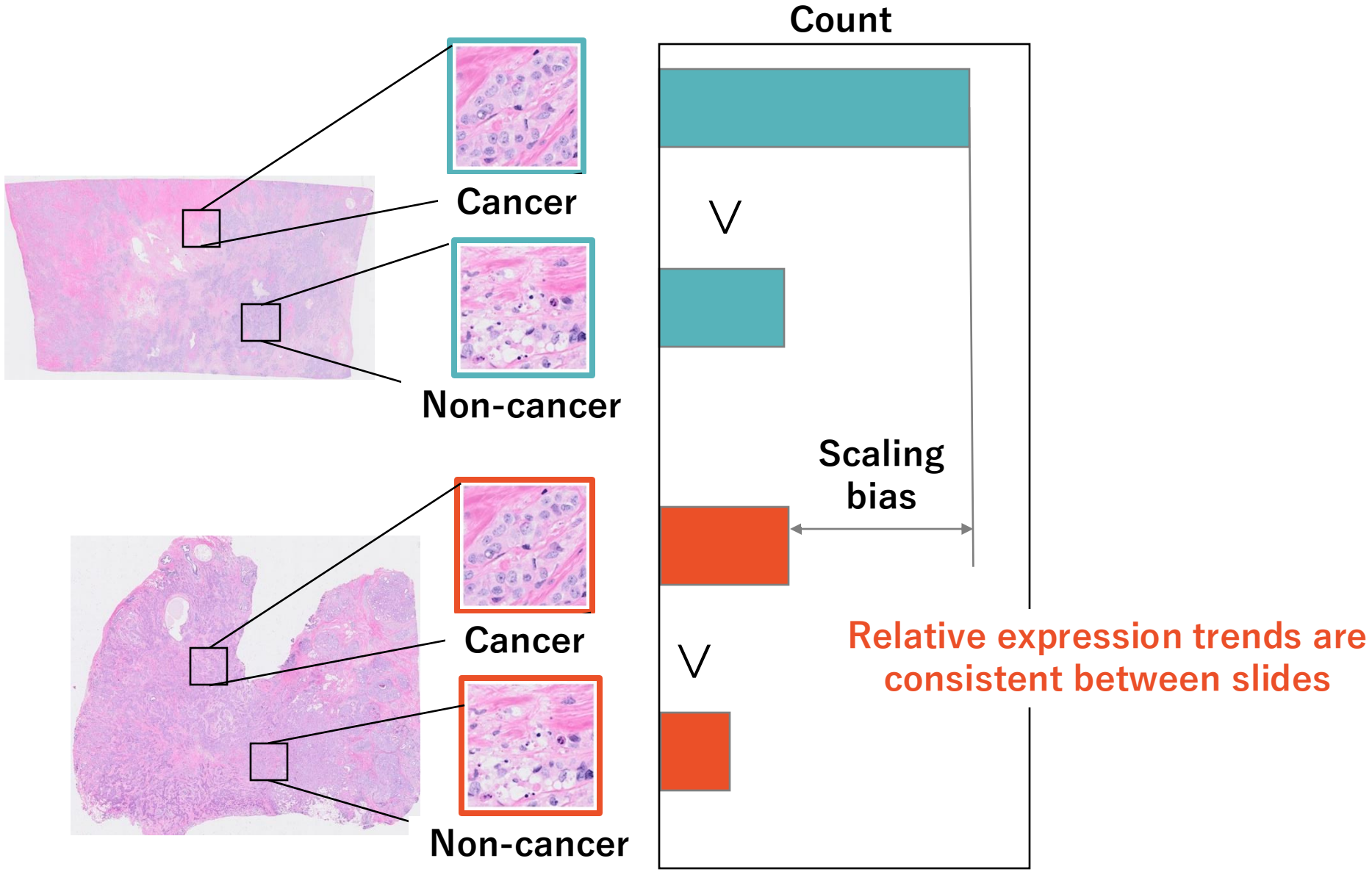


Stochastic noise

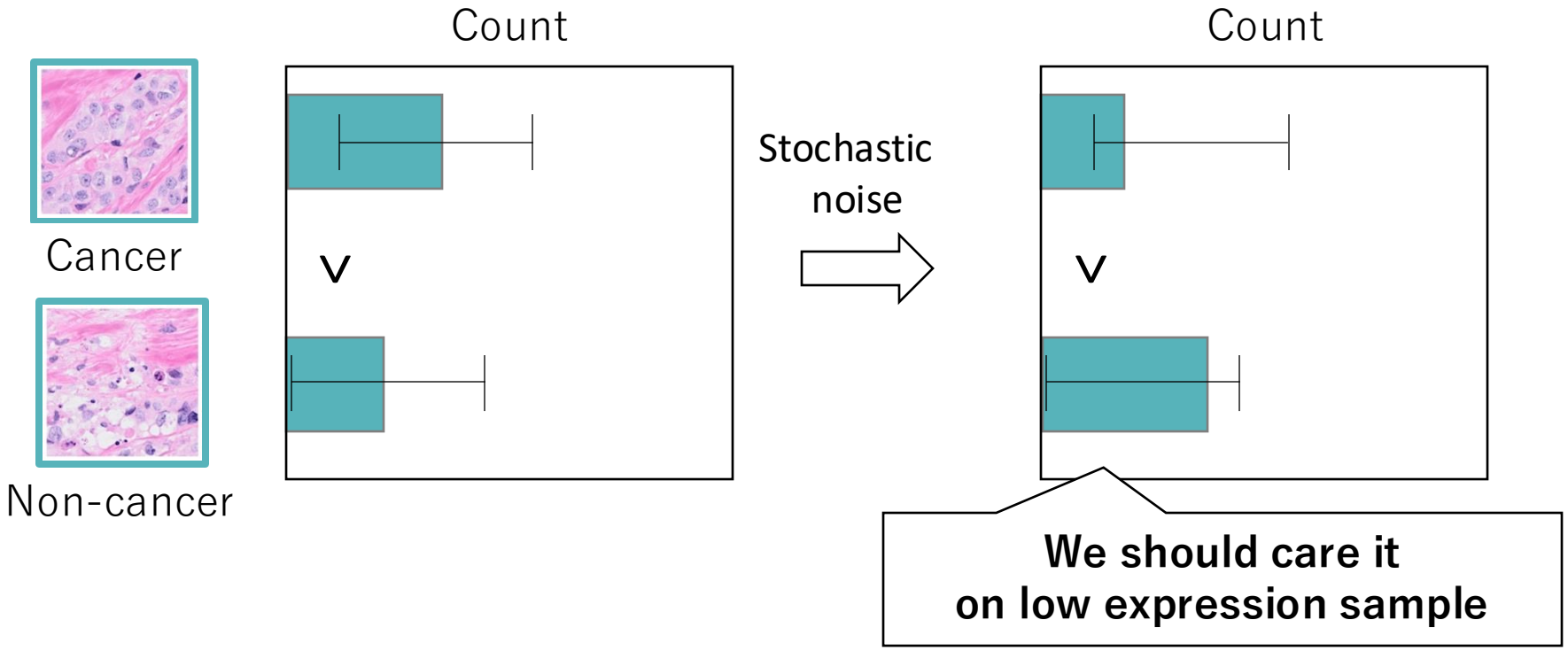
- Temporal dynamics
- Heterogeneity of cell
- Follows count dist.



Learning to relative expression trends

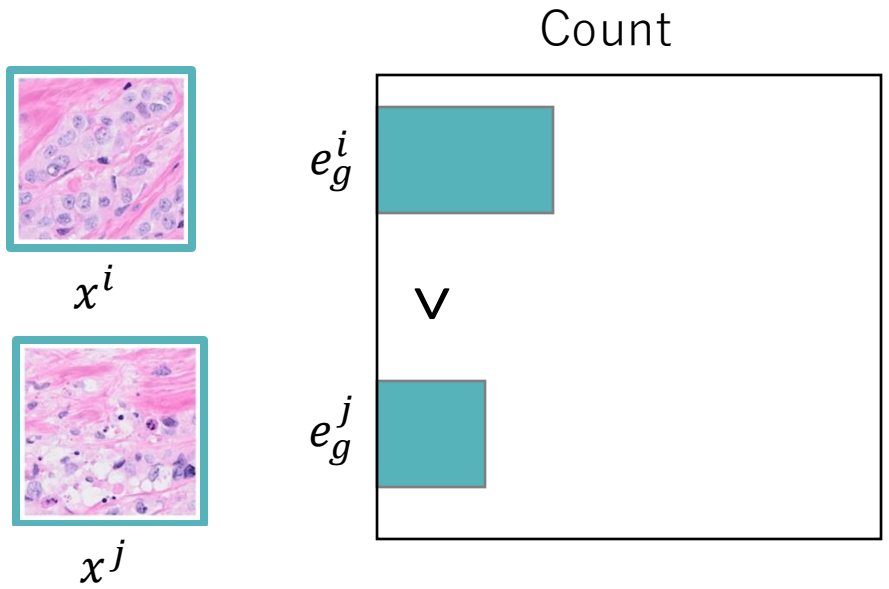


Stochastic noise may flip relation



Models gene expression as a discrete probabilistic distribution!

Modeling expression given pair or list



$$P\left(e^i \middle| x^i, x^j, e_g^i + e_g^j\right) = \prod_{g=1}^{N_g} P\left(e_g^i \middle| x^i, x^j, e_g^i + e_g^j\right)$$

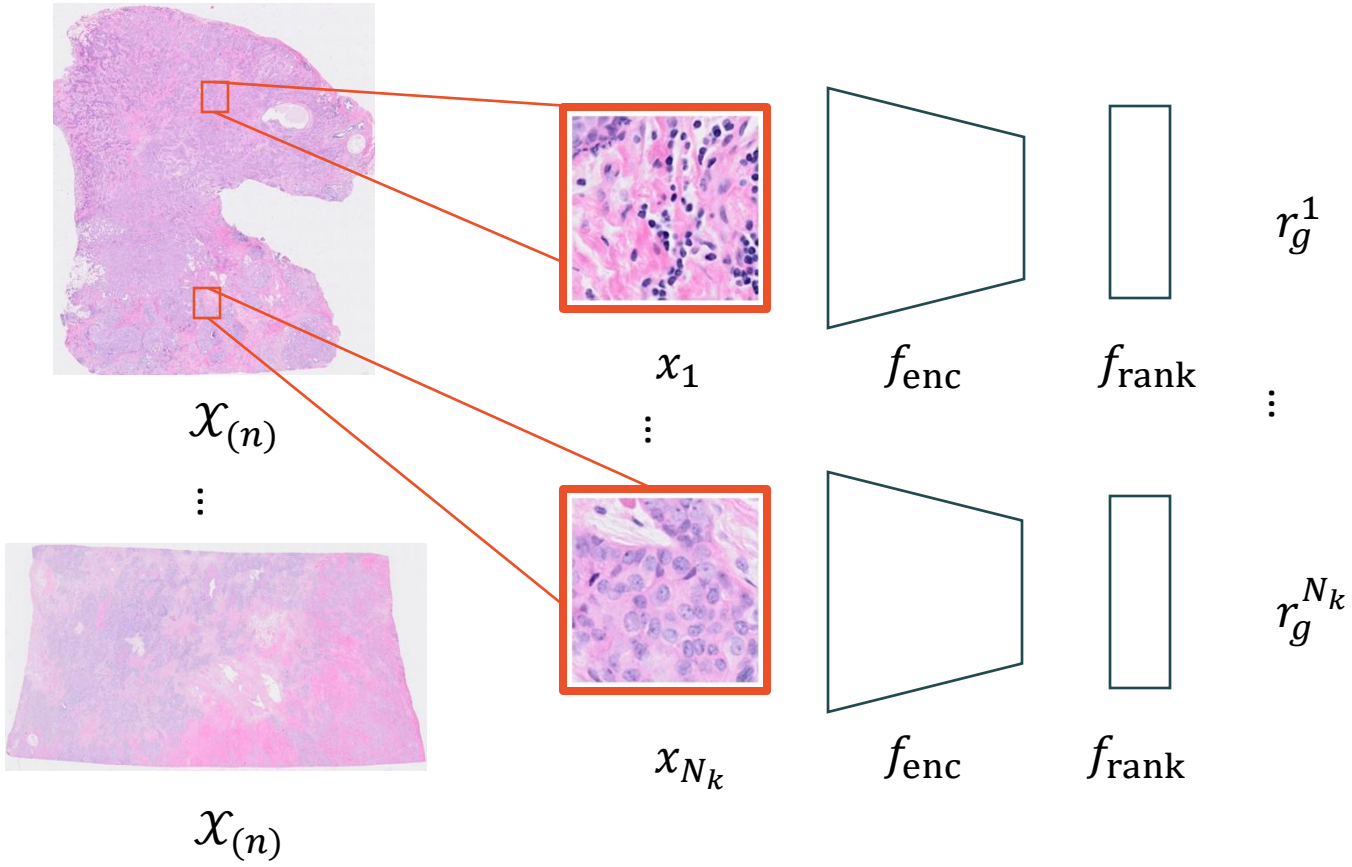
$$P\left(e_g^i \middle| x^i, x^j, e_g^i + e_g^j\right) = \text{Binomial}\left(\mathrm{e}_g^i + \mathrm{e}_g^j, \mathrm{p}_g^{i|j}\right)$$

Assume Binomial distribution by
Viewed as consequence of the counting process given relative frequencies

Learning to relative expression with modeling

Aim: estimate rank score r_g^i which reflect expression relation

If $e_g^i > e_g^j$, r_g^i should $r_g^i > r_g^j$

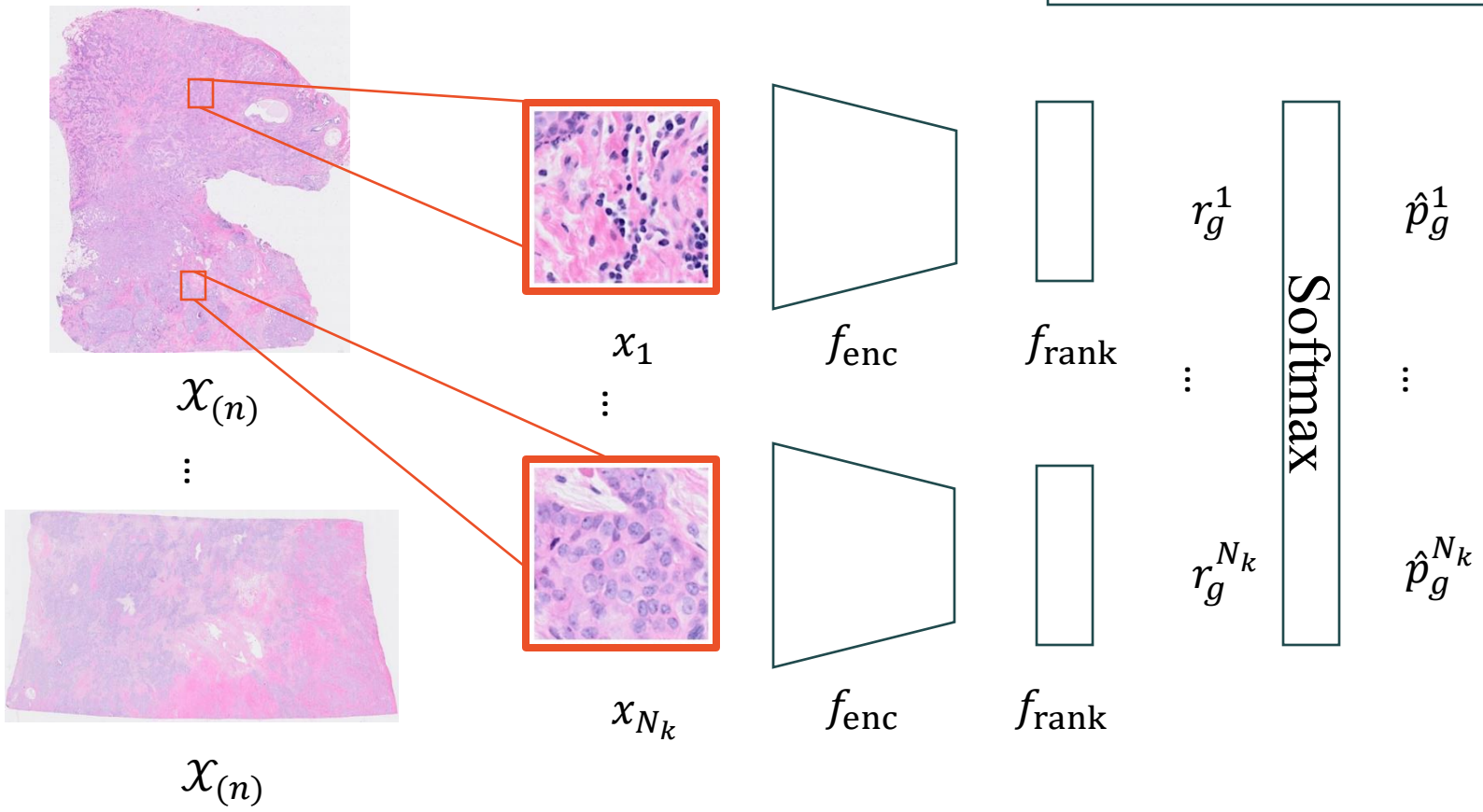


Learning to relative expression with modeling

$$\text{Loss} = -\log P\left(e^i \middle| x^i, x^j, e_g^i + e_g^j\right) = \sum_{g=1}^{N_g} P\left(e_g^i \middle| x^i, x^j, e_g^i + e_g^j\right)$$

$$= -\sum_{g=1}^{N_g} \left(e_g^i \log \hat{p}_g^i + e_g^j \log \hat{p}_g^j\right)$$

$$P\left(e_g^i \middle| x^i, x^j, e_g^i + e_g^j\right) = \text{Binomial}\left(e_g^i + e_g^j, \hat{p}_g^i\right)$$



Experiments on synthetic 1d data to understand characteristic of loss function

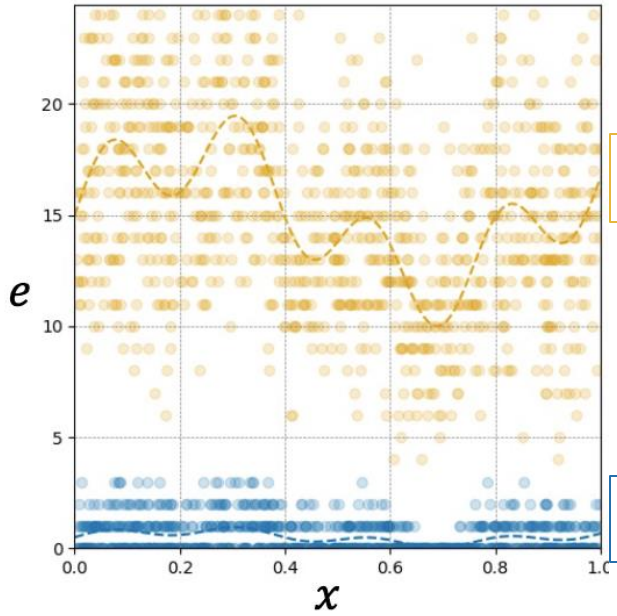
Gene expression follows negative binomial distribution

$$e^{(i)} \sim \mathrm{NB}\left(\mu\left(x^{(i)}\right), \mathrm{r}\right)$$

$$\mathfrak{D} = \left\{x^{(i)}, e^{(i)}\right\}$$

Due to batch effect the signal will be change:

Scaling factor $\alpha\ \mu\left(x^{(i)}\right) + \beta$ additive factor



Tissue 2: $e^{(i)} \sim \mathrm{NB}(10\mu\left(x^{(i)}\right) + 10, \mathrm{r})$ (α = 10, β = 10)

Tissue 1: $e^{(i)} \sim \mathrm{NB}\left(\mu\left(x^{(i)}\right), \mathrm{r}\right)$

Experiments on synthetic 1d data to understand characteristic of loss function

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

Not consider relative expression

Not consider noise

Experiments on real dataset (HEST 1k)

Evaluated on 7 types of organ

	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	*0.485*	0.336	0.524	**0.172**	0.091	0.134	0.293
	NB	0.199	**0.491**	0.119	0.538	*0.160*	0.075	0.126	0.244
Pair	Rank	0.317	0.317	0.181	0.566	0.047	0.059	0.110	0.228
	PairSTrank	*0.494*	0.458	**0.346**	*0.613*	0.136	**0.127**	*0.228*	*0.343*
List	PCC	0.472	0.459	0.307	**0.640**	0.105	0.102	0.198	0.326
	ListSTrank	**0.510**	0.459	*0.343*	0.597	0.140	*0.125*	**0.238**	**0.345**

Summary

Task. Patch-level gene expression estimation

Difficulty. batch effect and stochastic noise of observed data

Idea.

1. Batch effect
   ->learning to relative expression
2. Stochastic noise
   -> Model relative expression with discrete distribution