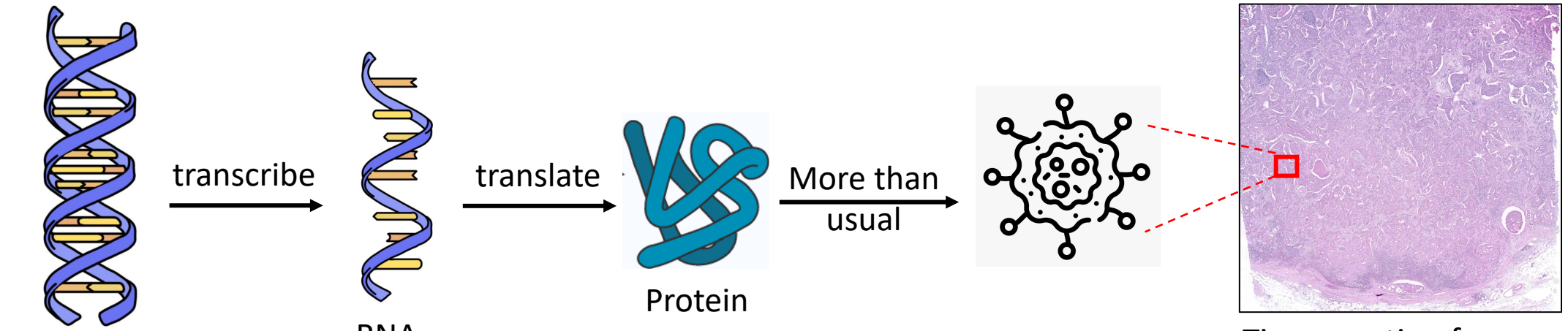
Predicting Spatial Transcriptomics from Histology Images via Biologically Informed Flow Matching

Tinglin Huang

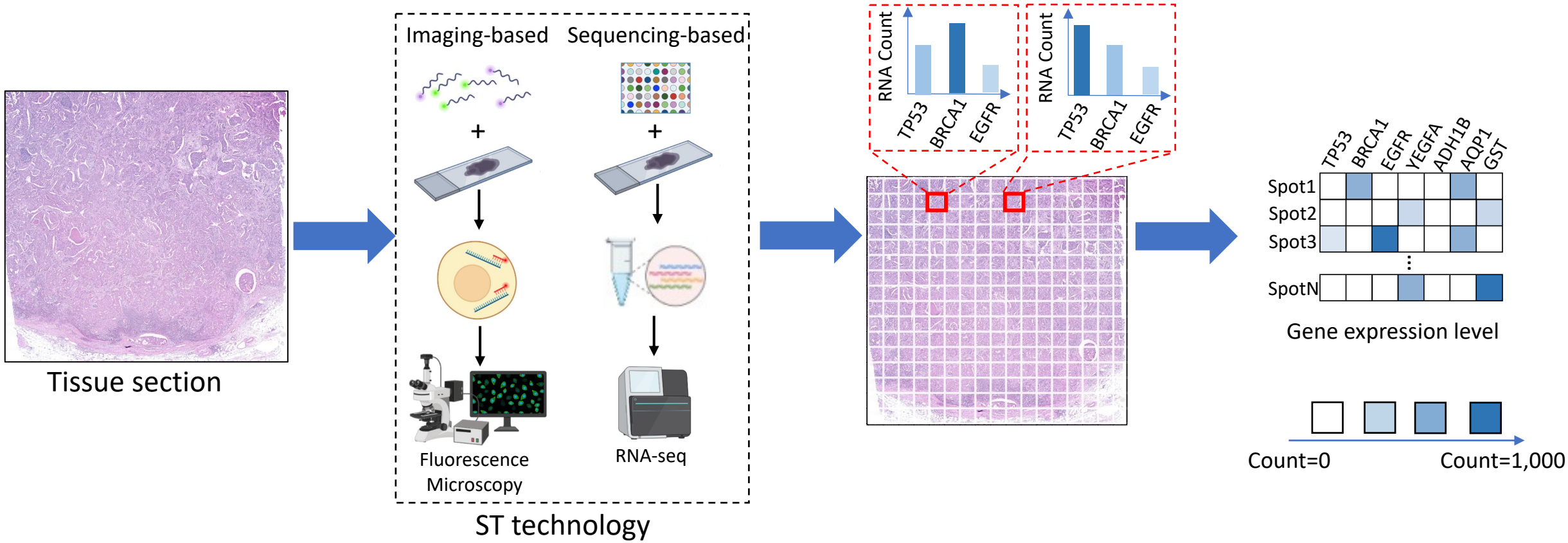
Background

- Gene expression regulates the cellular processes by central dogma
  - E.g., the overexpression of MYC gene can lead to cancer cells
  - Detecting the gene expression level can monitor the disease progression and treatment response



Spatial Transcriptomics (1)

- A technology that can detect spatial distribution of RNA transcripts
  - The number of RNA transcripts represents the gene expression level

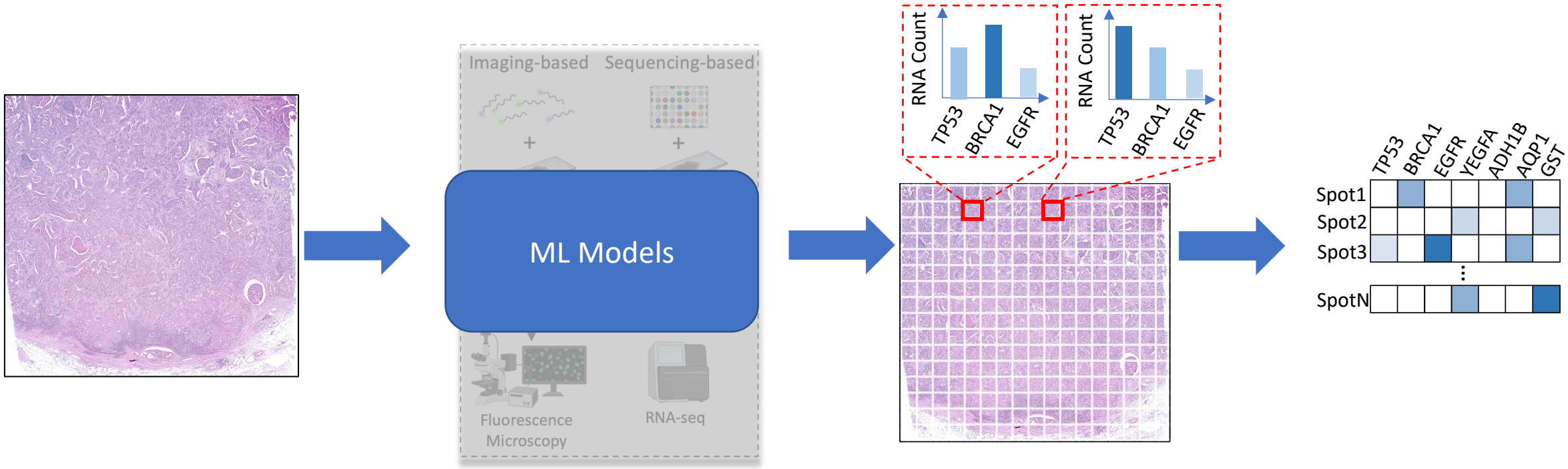


Spatial Transcriptomics (2)

- ST methods are low throughput and rely on specialized equipment
  - A single slide costs approximately $1,000-1,500 USD and 7-10 days

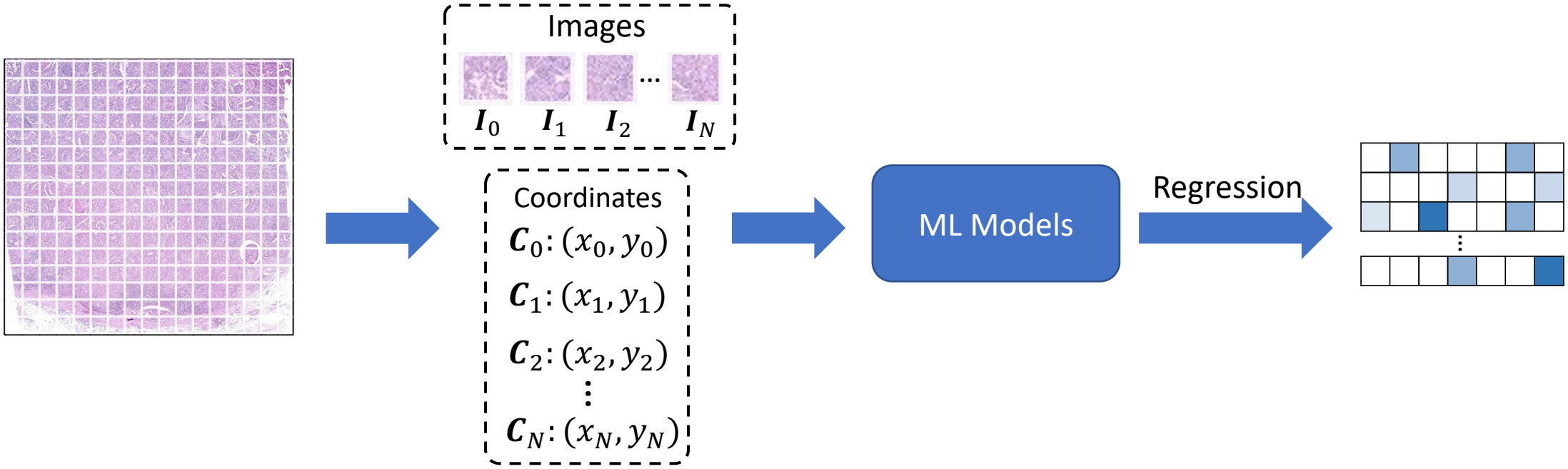
Spatial Transcriptomics (3)

- ML is one of the promising alternatives



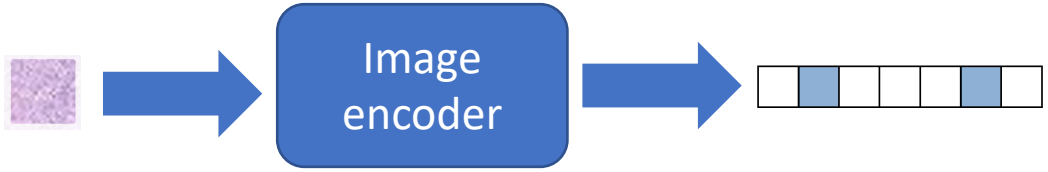
Problem formulation

- Predict the gene expression for each spot (regression task)
  - Input: spot images $\{I_1, \cdots, I_N\}$, coordinates $\{C_1, \cdots, C_N\}$
  - Output: gene expression $\{Y_1, \cdots, Y_N\}$

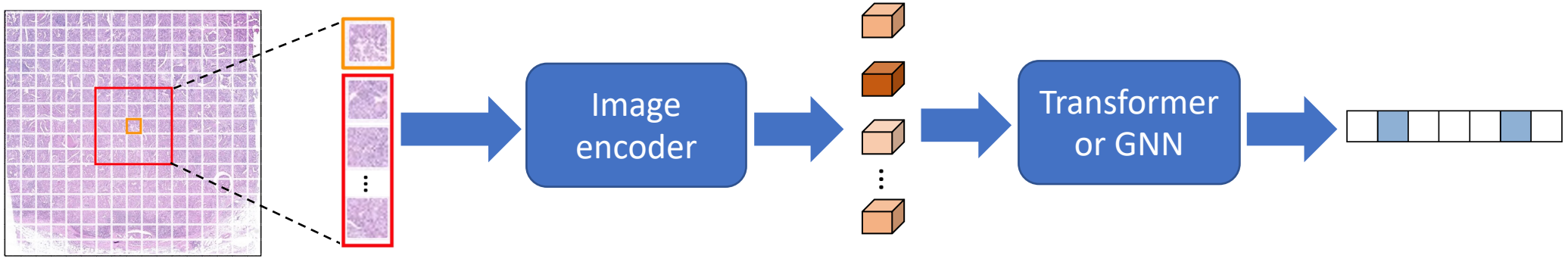


Modeling strategies

- Spot-level methods solely utilize images: $p(Y_i|I_i)$
  - Pathology foundation model shows great performance

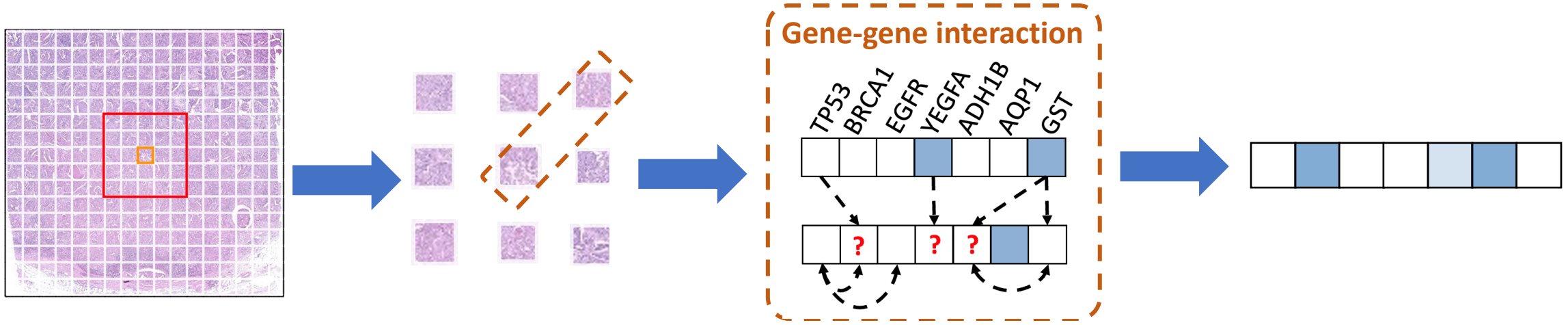


- Slide-level methods incorporate the slide context: $p(Y_i|I_1, \cdots, I_N)$
  - Aggregate the neighboring spots with Transformer/GNN



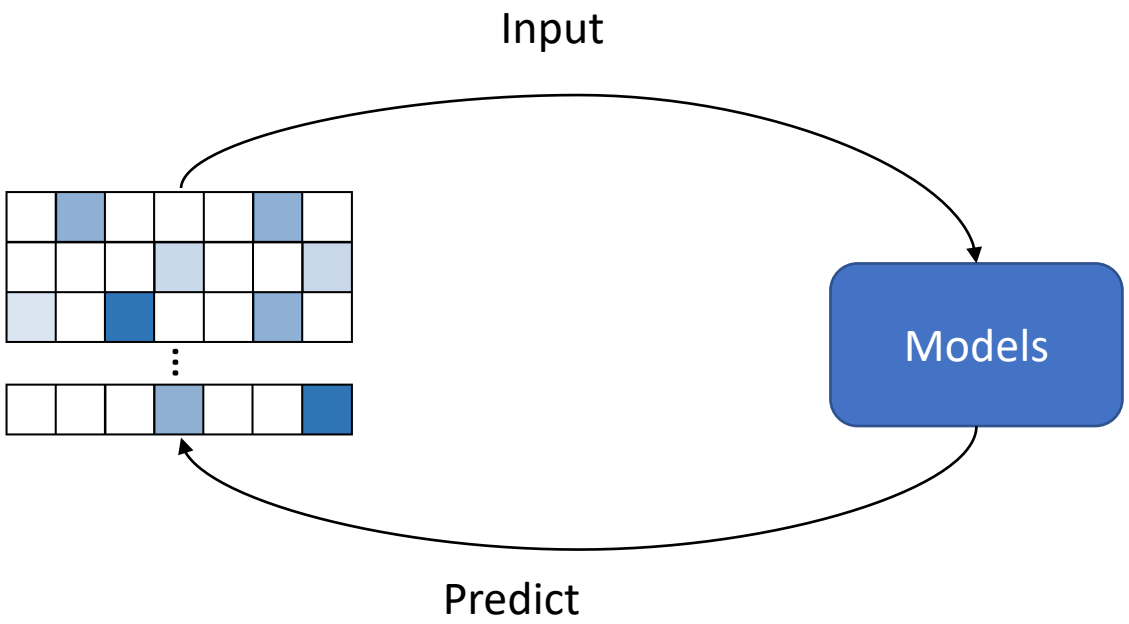
Gene-Gene Interaction

- The interaction between genes regulates the expression significantly
  - Gene regulation pathway: gene1→gene2→gene3
  - But it has been overlooked so far



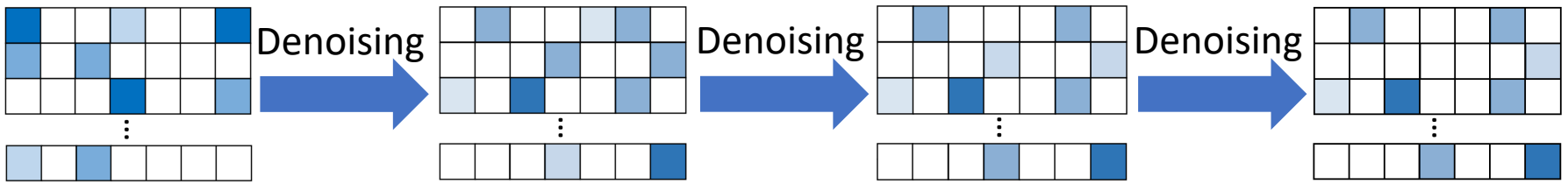
Chicken-and-Egg Problem

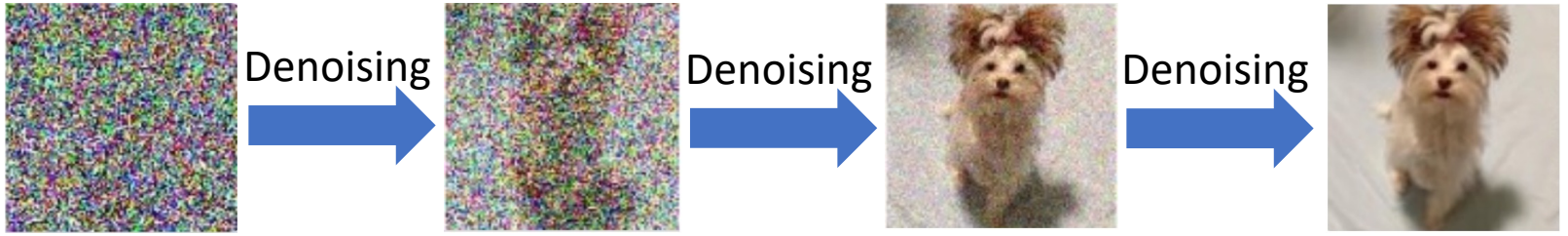
- Gene-gene interaction requires gene expression as context, yet gene expression itself is the target of prediction



Generative Gene Expression Prediction

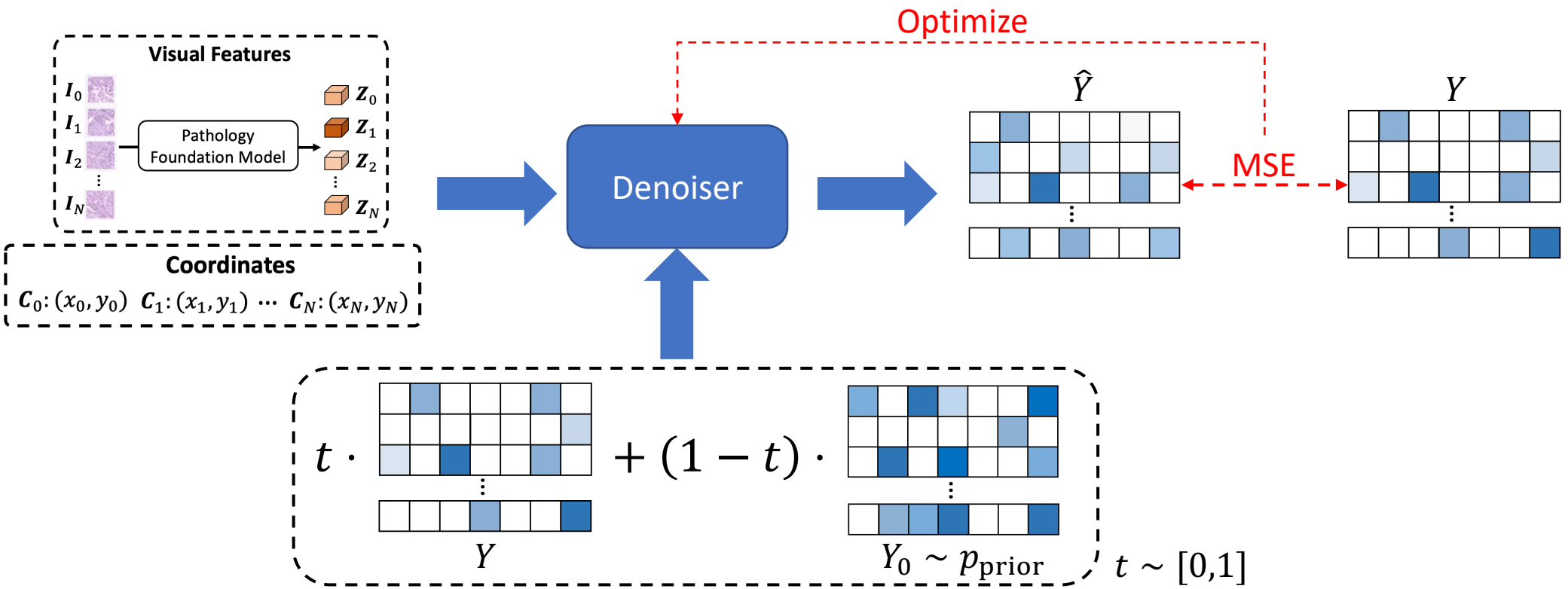
- Generative model provides a recipe to receive gene expression as input
  - Iteratively refine the gene expression instead of one-step regression





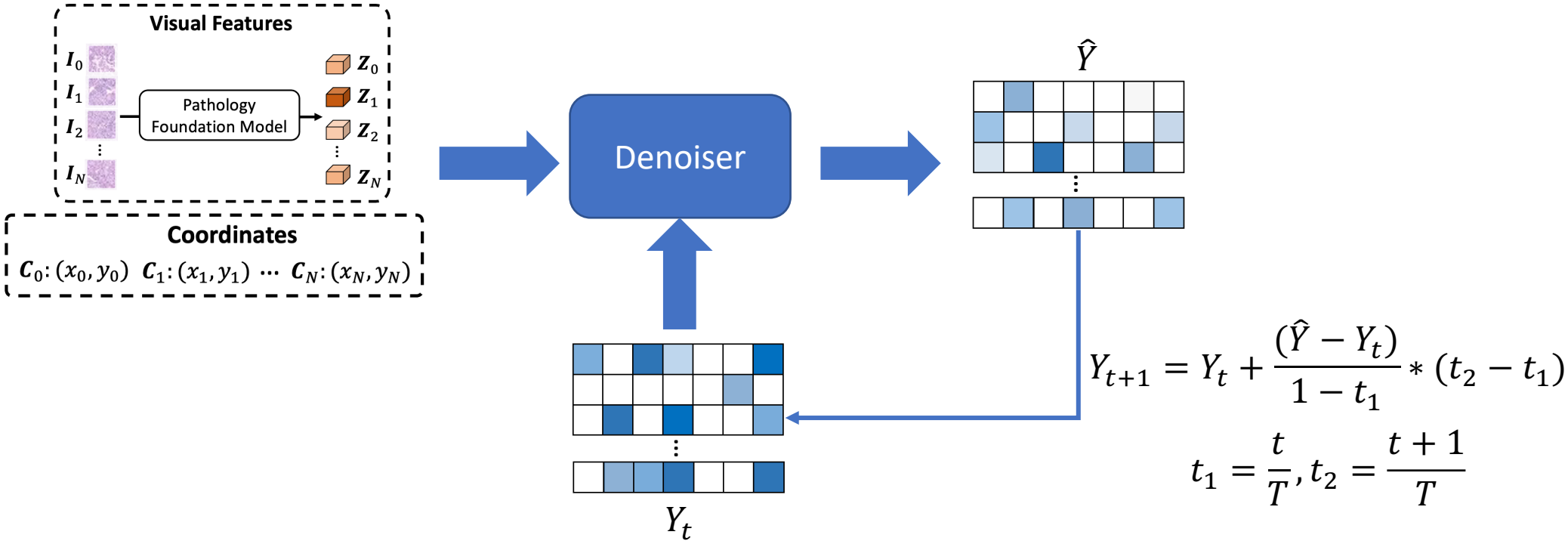
Flow Matching: Train

- Optimized by recovering gene expression from the interpolation between ground truth and a sampled noise



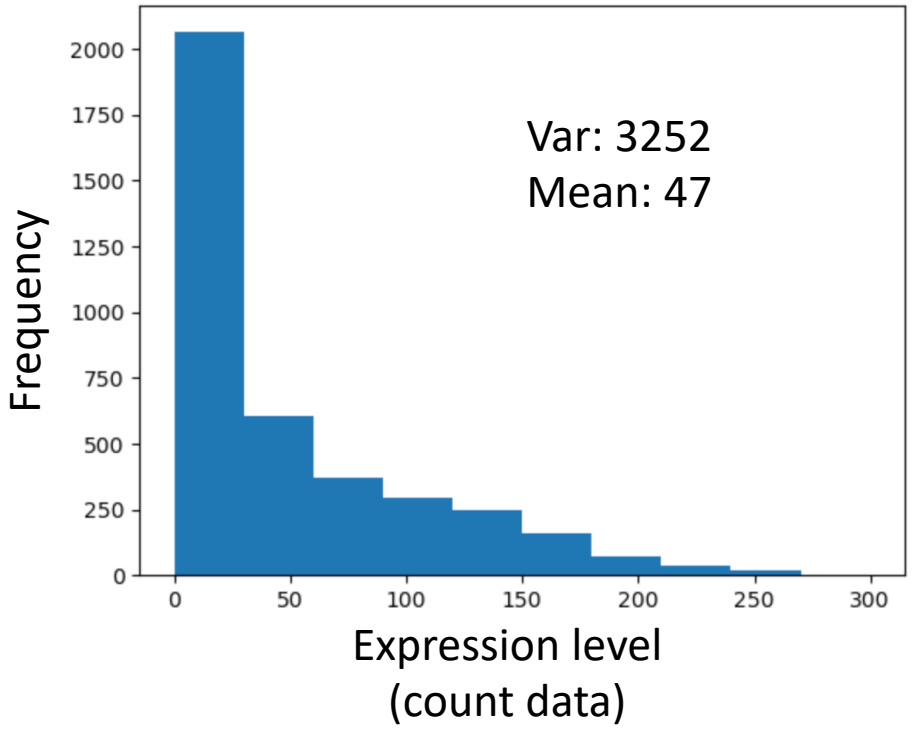
Flow Matching: Inference

- Iteratively refines the gene expression by interpolating between the noisy input Y_t and the predicted expression $\hat{Y}$



ZINB distribution as prior

- Gene expression is typically sparse and exhibits overdispersion
  - Overdispersion: variance > mean



- Negative binomial distribution is suitable for representing overdispersed data
- It is further extended to **zero-inflated negative binomial** (ZINB) distribution for modeling the sparse characteristic
  - The sample can randomly be dropped

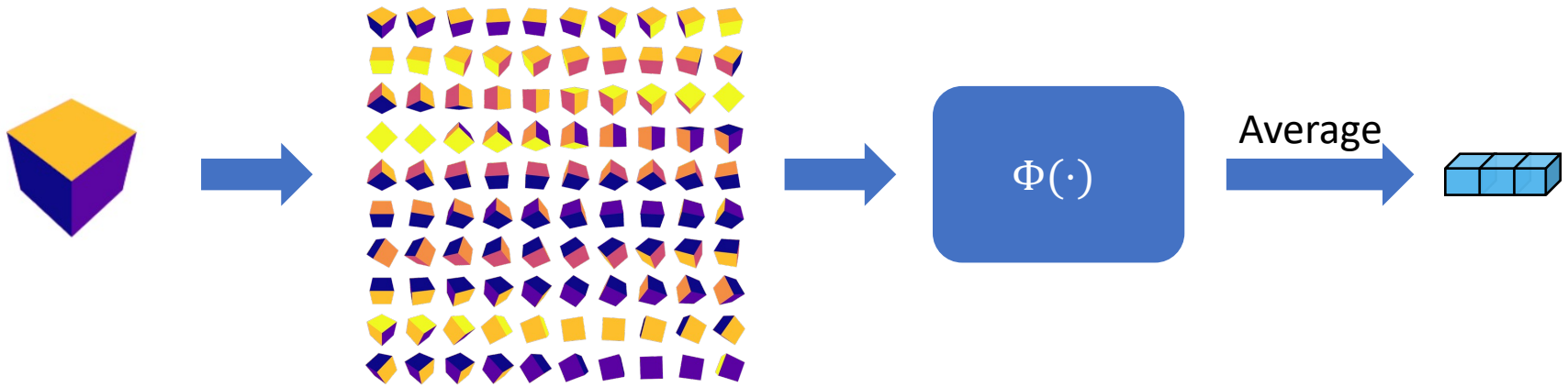
$$p(y \mid \mu, \phi, \pi) = \begin{cases} \pi + (1-\pi)\left(\frac{\Gamma(y+\phi)}{\Gamma(\phi)\, y!}\right)\left(\frac{\phi}{\phi+\mu}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{y} & \text{if } y = 0, \\ (1-\pi)\left(\frac{\Gamma(y+\phi)}{\Gamma(\phi)\, y!}\right)\left(\frac{\phi}{\phi+\mu}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{y} & \text{if } y > 0, \end{cases}$$

Negative binomial distribution

y: count, π: dropout probability, μ: mean, ϕ: number of failures

Architecture of Denoiser (1)

- Input: coordinates, visual features, gene expression
  - $C \in \mathbb{R}^{N\times 2}, Z \in \mathbb{R}^{N\times D}, Y \in \mathbb{R}^{N\times G}$
- Extend Frame Averaging Transformer[1]
  - Goal: invariantly embed the spot with rotation/translation on coordinates
  - Any model $\Phi(\cdot)$ is invariant if it considers all kinds of transformation t



[1] Protein-Nucleic Acid Complex Modeling with Frame Averaging Transformer, NeurIPS2024

Architecture of Denoiser (2)

- Frame averaging (FA) prove the group average can be conducted over **a carefully selected subset** $\mathcal{F}(X) \subset G_{\text{all}}$
  - Frame $\mathcal{F}$ is a mapping from coordinate space into a group
- Such a subset can be calculated by PCA
  - Two principal components $\vec{u}_1, \vec{u}_2$
  - $\vec{u}_i \in \mathbb{R}^{2\times 1}$

$C \in \mathbb{R}^{N\times 2}$ → PCA → $\vec{u}_1, \vec{u}_2$ → $\{\pm\vec{u}_1, \pm\vec{u}_2\}$

Totally 2^2=4 rotation matrix

Each matrix $\in \mathbb{R}^{2\times 2}$

Architecture of Denoiser (3)

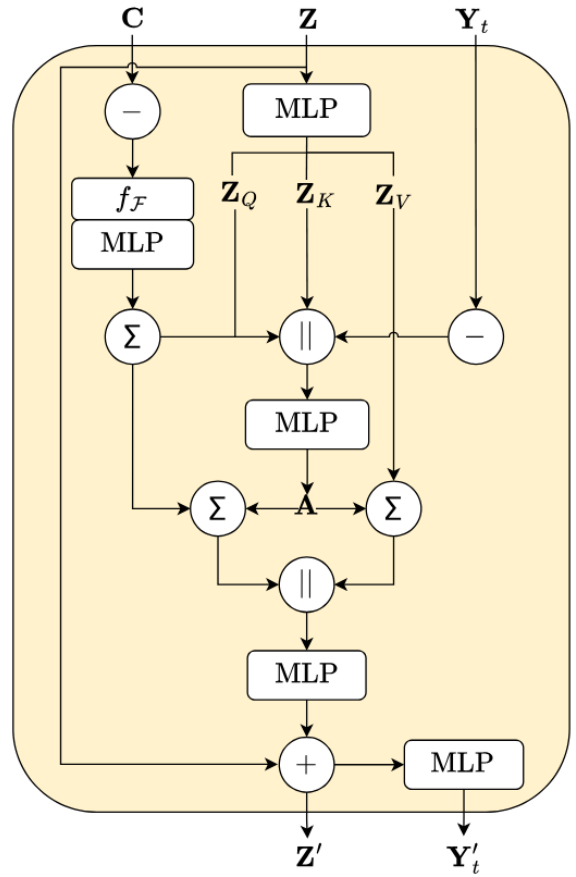
- Extend FAFormer[1] to incorporate all these information
  - $\mathcal{N}(i)$: k-nearest neighbors of i-th spot
  - Coordinate is embedded with frame averaging

$$\left\{c'_{i\to j}\right\}_{j\in\mathcal{N}(i)} = \mathrm{FAMLP}(c_i, \left\{c_j\right\}_{j\in\mathcal{N}(i)})$$

  - Attention and aggregation

$$a_{ij} = \mathrm{Softmax}\left(\mathrm{MLP}\left(z_i^Q || z_i^K || c'_{i\to j} || (y_i - y_j)\right)\right)$$

$$z = \mathrm{MLP}\left(\sum_{j\in\mathcal{N}(i)} a_{ij} z_j^V \,||\, \sum_{j\in\mathcal{N}(i)} a_{ij} c'_{i\to j}\right) + z$$



[1] Protein-Nucleic Acid Complex Modeling with Frame Averaging Transformer, NeurIPS2024

Benchmark

- 10 datasets covering 48 patients and 74 slides
  - Metric: Pearson correlation between prediction and groundtruth
- Baselines
  - Pathology foundation models: Ciga, UNI, Gigapath
  - Slide-level approaches: STNet, BLEEP, Hist2ST, HisToGene, TRIPLEX

	Spot-based					**Slide-based**						
	Ciga	UNI	Gigapath	STNet DenseNet121	BLEEP ResNet50	Gigapath-slide	Hist2ST ViT	HisToGene ViT	TRIPLEX Ciga	Ciga	STFlow UNI	Gigapath
IDC	$0.423_{.002}$	$0.502_{.050}$	$0.514_{.064}$	$0.380_{.048}$	$0.346_{.094}$	OOM	$0.052_{.032}$	$0.350_{.063}$	$0.492_{.042}$	$0.460_{.028}$	$0.589_{.063}$	$0.565_{.055}$
PRAD	$0.343_{.001}$	$0.357_{.000}$	$0.386_{.008}$	$0.346_{.006}$	$0.303_{.004}$	$0.386_{.006}$	$0.065_{.038}$	$0.253_{.005}$	$0.351_{.023}$	$0.380_{.001}$	$0.420_{.005}$	$0.415_{.013}$
PAAD	$0.406_{.008}$	$0.424_{.060}$	$0.436_{.054}$	$0.370_{.047}$	$0.347_{.059}$	$0.394_{.041}$	$0.111_{.004}$	$0.303_{.007}$	$0.429_{.045}$	$0.440_{.047}$	$0.506_{.078}$	$0.513_{.063}$
SKCM	$0.492_{.003}$	$0.613_{.020}$	$0.578_{.001}$	$0.385_{.054}$	$0.407_{.130}$	$0.543_{.014}$	$0.195_{.010}$	$0.321_{.028}$	$0.576_{.091}$	$0.608_{.072}$	$0.707_{.028}$	$0.651_{.089}$
COAD	$0.275_{.054}$	$0.287_{.005}$	$0.287_{.008}$	$0.249_{.063}$	$0.172_{.014}$	OOM	$0.071_{.006}$	$0.266_{.015}$	$0.305_{.004}$	$0.344_{.023}$	$0.328_{.013}$	$0.325_{.023}$
READ	$0.051_{.005}$	$0.162_{.080}$	$0.151_{.081}$	$0.116_{.032}$	$0.098_{.063}$	$0.188_{.048}$	$0.034_{.025}$	$-0.006_{.013}$	$0.129_{.062}$	$0.137_{.075}$	$0.243_{.002}$	$0.260_{.023}$
CCRCC	$0.136_{.005}$	$0.186_{.050}$	$0.187_{.062}$	$0.213_{.071}$	$0.107_{.023}$	$0.183_{.052}$	$0.100_{.053}$	$0.112_{.036}$	$0.229_{.036}$	$0.250_{.054}$	$0.335_{.070}$	$0.326_{.065}$
HCC	$0.042_{.001}$	$0.051_{.000}$	$0.054_{.002}$	$0.078_{.034}$	$0.066_{.021}$	$0.026_{.005}$	$0.015_{.001}$	$0.028_{.015}$	$0.044_{.022}$	$0.105_{.030}$	$0.128_{.017}$	$0.125_{.019}$
LUNG	$0.544_{.001}$	$0.511_{.030}$	$0.568_{.038}$	$0.526_{.025}$	$0.476_{.021}$	$0.530_{.025}$	$0.302_{.063}$	$0.477_{.057}$	$0.563_{.036}$	$0.584_{.027}$	$0.608_{.021}$	$0.602_{.013}$
LYMPH	$0.235_{.006}$	$0.234_{.050}$	$0.275_{.049}$	$0.237_{.063}$	$0.204_{.061}$	$0.284_{.042}$	$0.096_{.079}$	$0.238_{.062}$	$0.286_{.055}$	$0.307_{.052}$	$0.305_{.056}$	$0.305_{.053}$
Average	0.305	<u>0.347</u>	0.344	0.290	0.252	/	0.104	0.234	0.340	0.361	**0.419**	0.409

Summary

- Task: gene expression prediction based on histology image
- Repurpose one-step regression task as a generative task
  - Apply flow matching with a novel transformer architecture as denoiser
  - Introduce ZINB distribution as prior, which is biologically informed
- Achieve best performance over all the baselines on 10 datasets
  - STFlow is foundation model-agnostic, which can consistently achieve better performance