Bridging AI and BNP for Layered Point Pattern Data Analysis

NSF-ICERM

BNP Inference - Computational Issues

Qiwei Li

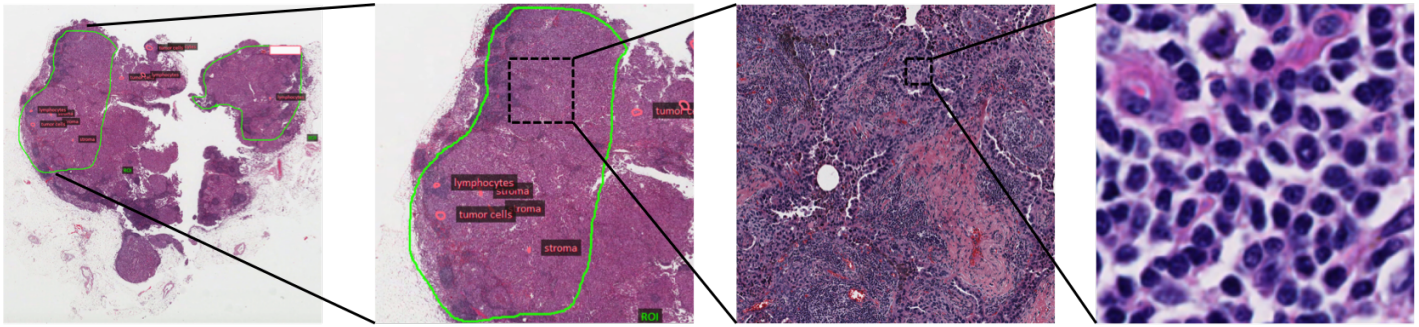
Associate Professor of Statistics

Department of Mathematical Sciences
The University of Texas at Dallas

Table of Contents

Tumor Pathology Images

- Capture histological details in high resolution
  - Average size: $25,000 \times 20,000 = 500$ megapixels



- Require a systematical exploration on subtle patterns
- Harbor a large amount of biological information
  - Cell growth pattern
    - Survival outcome (Gleason *et al.*, 2002; Amin *et al.*, 2002; Borczuk *et al.*, 2009; Barletta *et al.*, 2010)
    - Treatment response (Tsao *et al.*, 2015)
  - Cell-cell interaction
  - Cell interaction with the surrounding micro-environment

AI-Statistics for Pathology Image Analysis

- AI discovers complex patterns, but lacks interpretability

AI-Statistics for Pathology Image Analysis

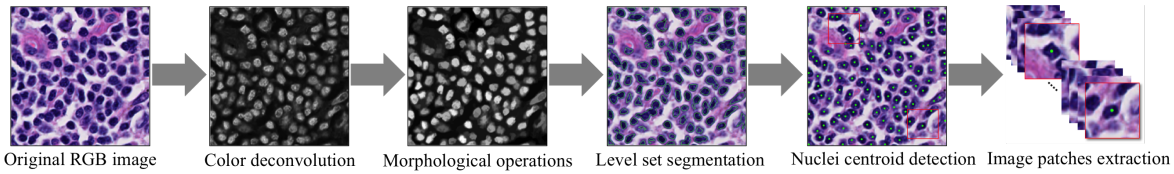
- AI discovers complex patterns, but lacks interpretability
- Statistics interprets data, but cannot afford complex data

AI-Statistics for Pathology Image Analysis

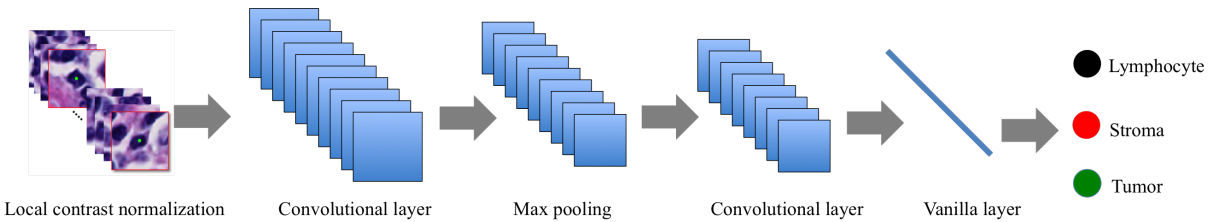
- AI discovers complex patterns, but lacks interpretability
- Statistics interprets data, but cannot afford complex data
- AI-Statistics workflow

AI - ConvPath Pipeline

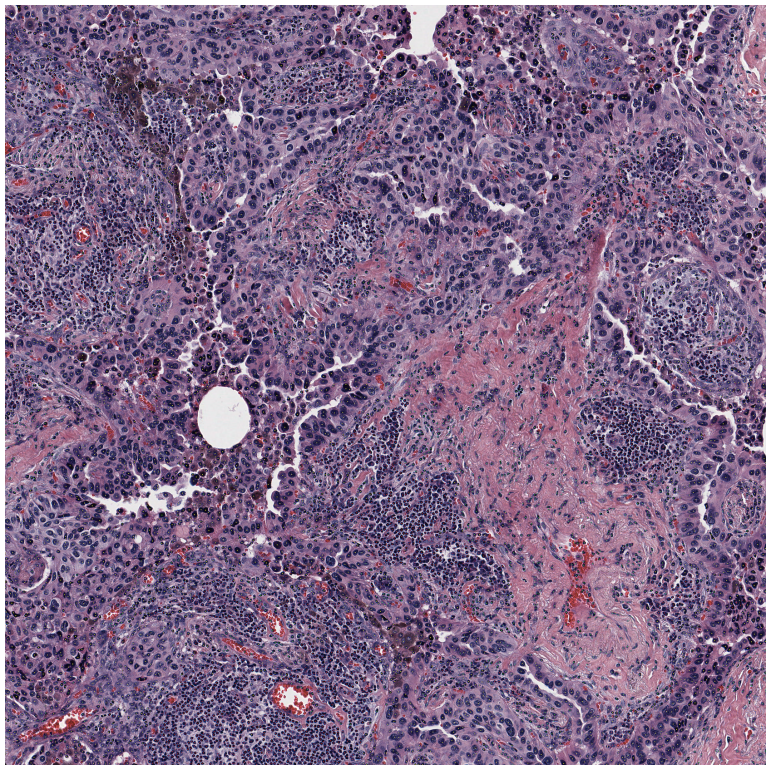
- Cell segmentation with water-shed method
  - Yi *et al.*, Automatic extraction of cell nuclei from H&E-stained histopathological images (2017), *J. Med. Imaging*, **4**(2)

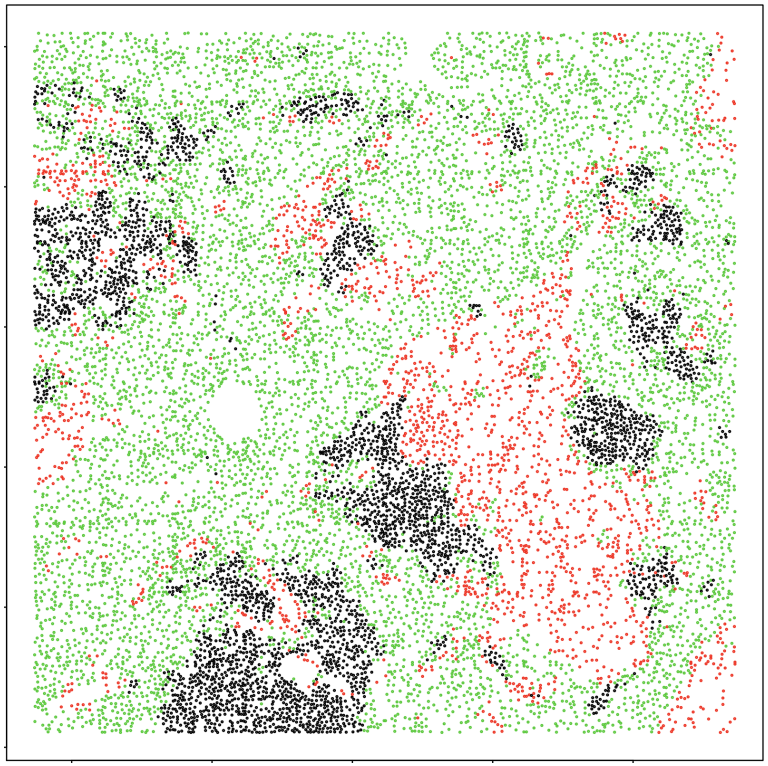


- Cell type classification by deep learning
  - Wang *et al.*, ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network (2019), *EBioMedicine*, **50**



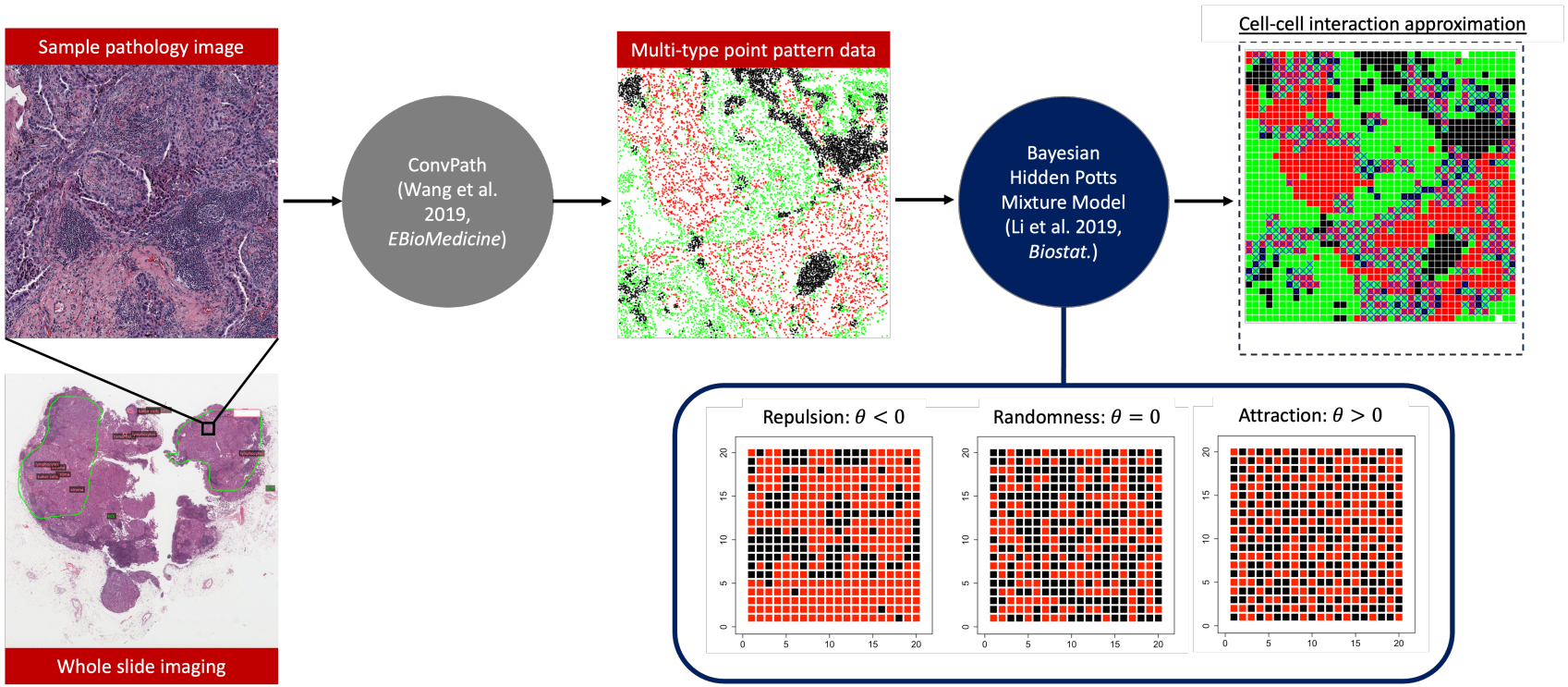
Data - Multi-type Point Patterns



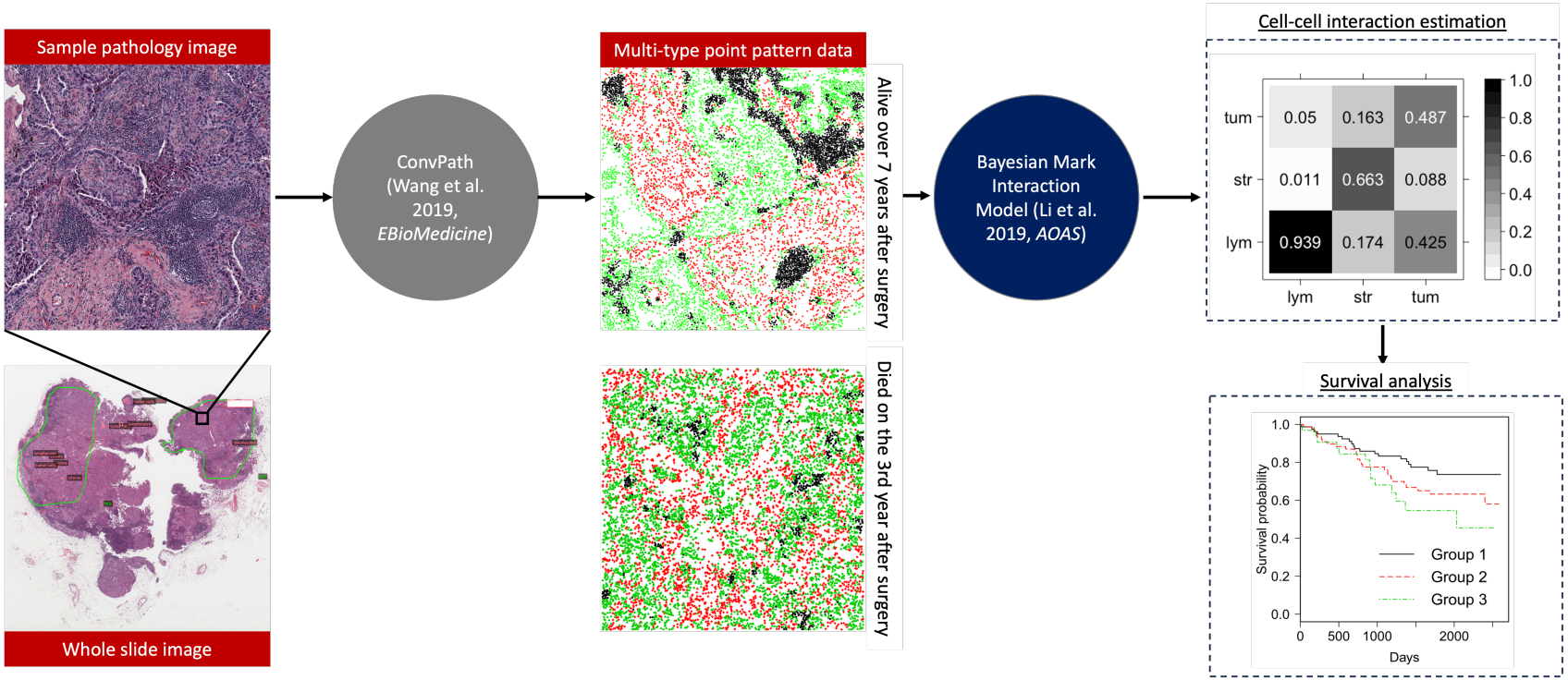


- Raw data: pathology images
- Processed data: multi-type point pattern data
  - Locations: $(x_i, y_i), i = 1, \ldots, n$
  - Marks: $z_i \in \{$Lymphocyte (●), stromal (●), and tumor (●)$\}$

AI-Bayesian for Point Pattern Data Analysis



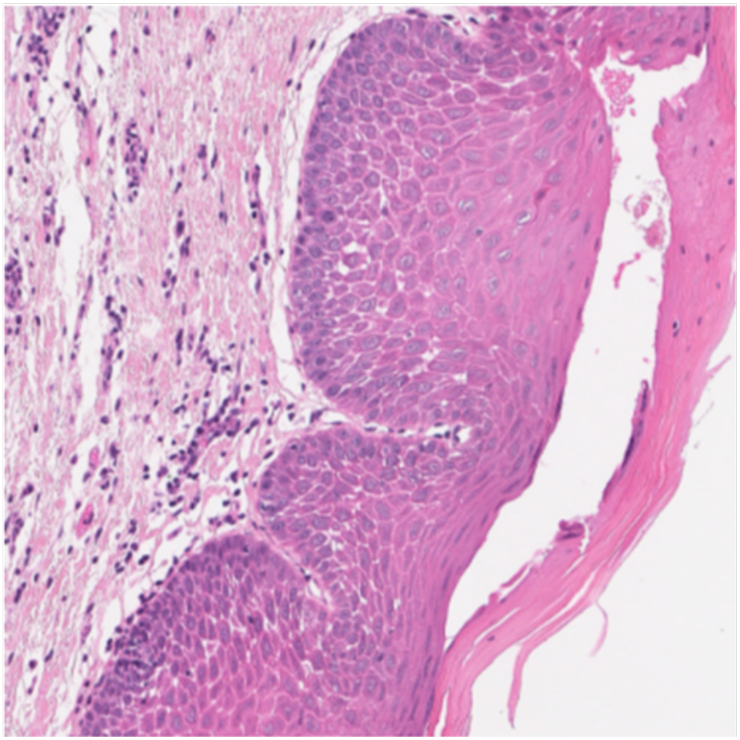
AI-Bayesian for Point Pattern Data Analysis



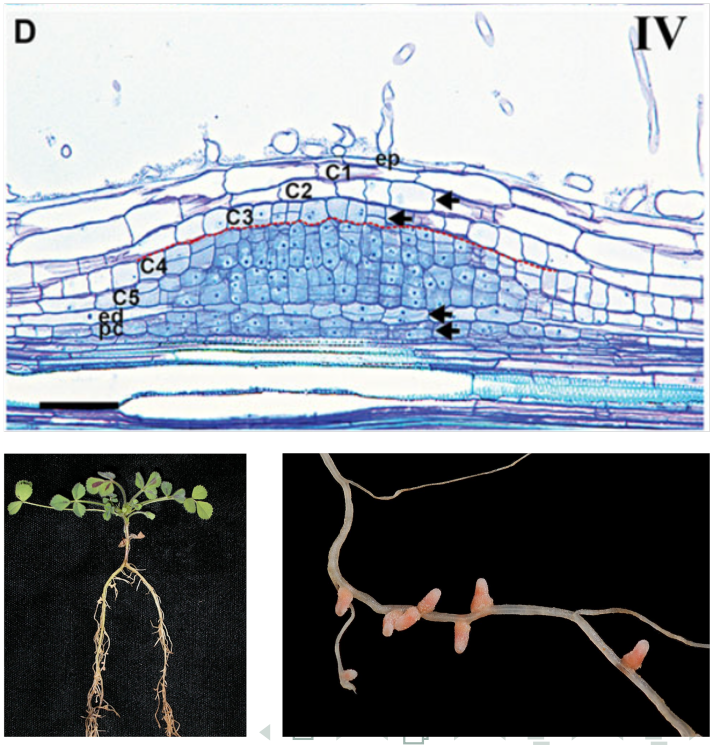
Motivation

- Motivation: the number of epithelial layers is linked to dysplasia severity in oral cancer
- Goal: estimate the number of cellular layers in an image

A cropped WSI from a patient in
the oral potentially malignant disorders (OPMD) study

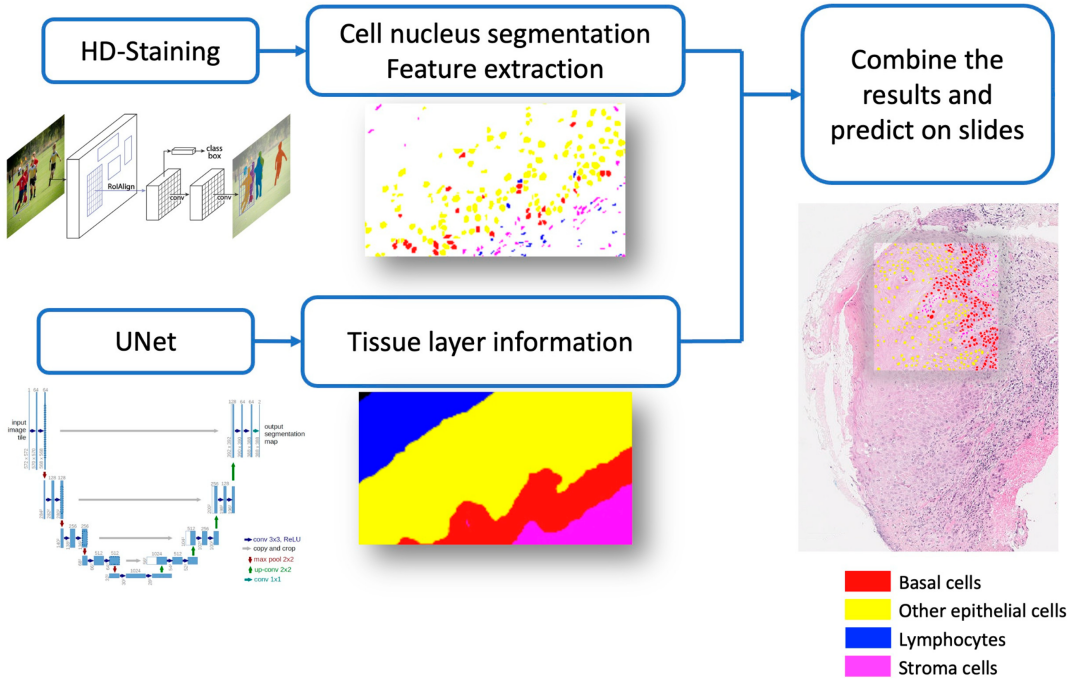


A micrograph image slide from *medicago truncatula* root nodule

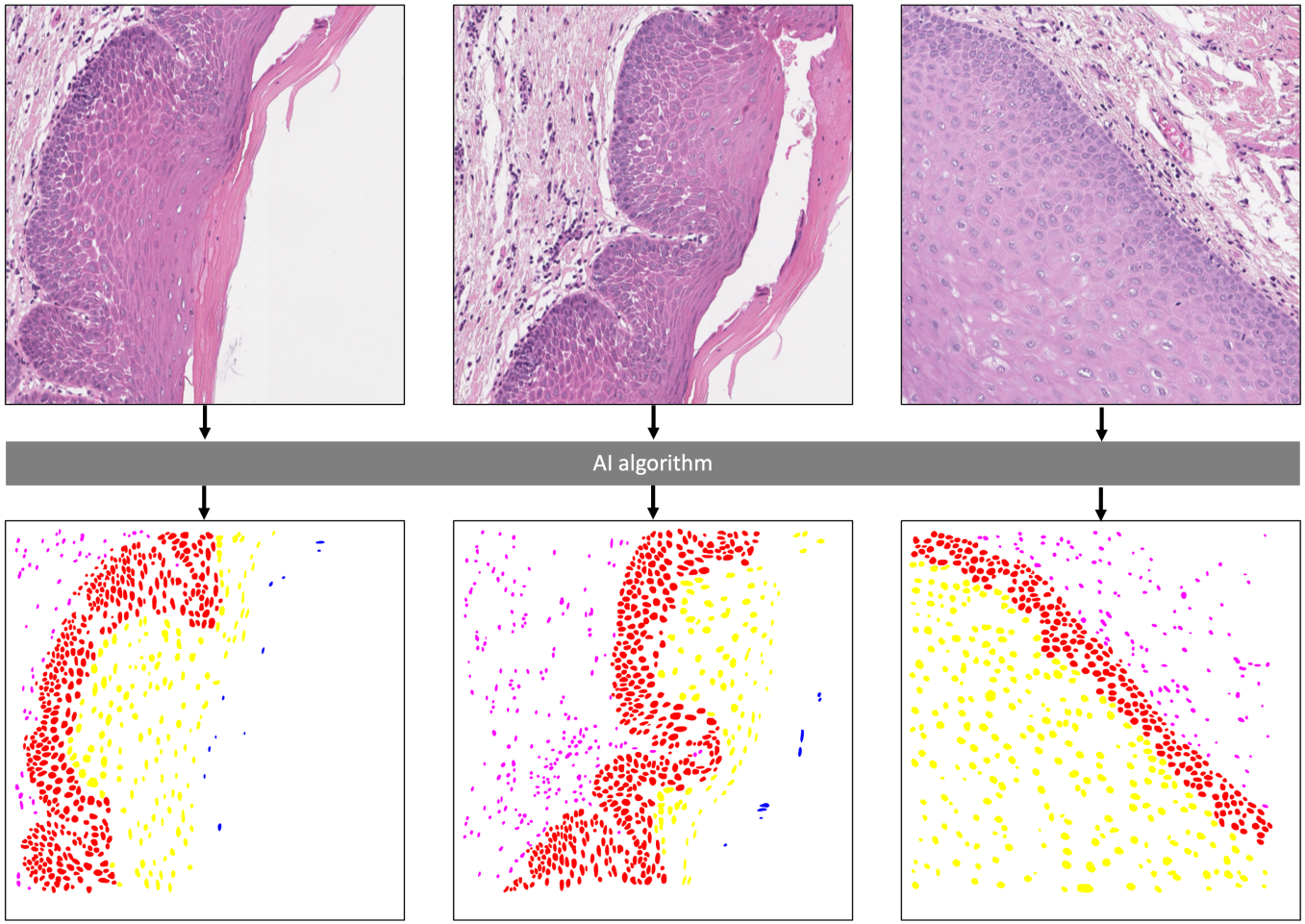


AI - Deep Convolutional Neural Network

- Cell nuclei segmentation and classification using HD-staining (Wang *et al.*, *Cancer Res.*, 2020) and YOLOv8 mask (Rong *et al.*, *Mod. Pathol.*, 2023)
- Tissue layer segmentation using U-Net (Ronneberger *et al.*, MICCAI'15)



Data - Multi-type Point Patterns

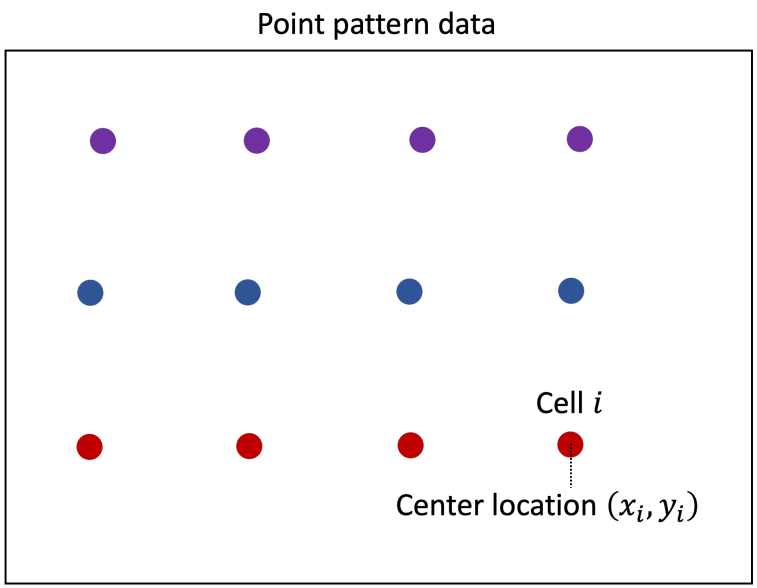


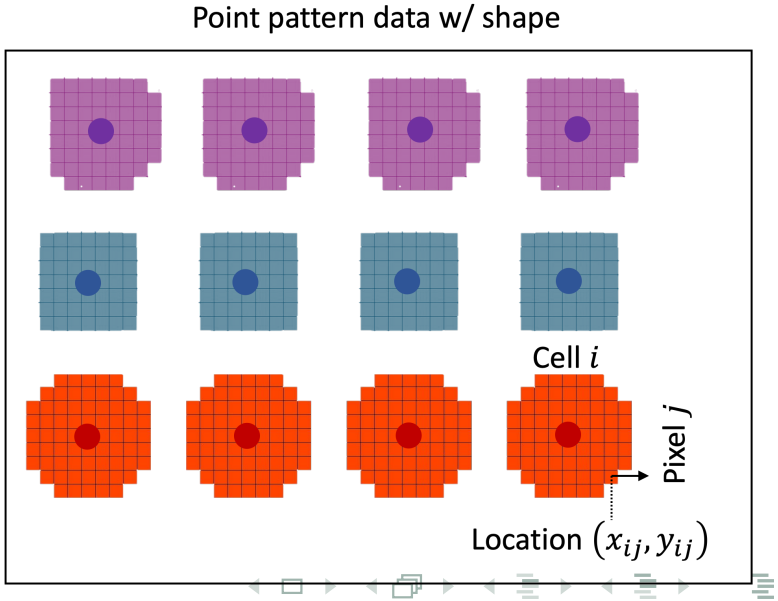
Data - Multi-type Point Patterns

- Raw data: oral tissue pathology images
  - 128 oral potentially malignant disorders (OPMD) patients from the Erlotinib Prevention of Oral Cancer (EPOC) trial at UT MDACC
  - 701 sample images (in 4 megapixels) from 255 whole slide images (WSI)
- Processed data: multi-type point pattern data
  - The number of cells n ranges from 169 to $3,003$
  - Locations: $(x_i, y_i), i = 1, \ldots, n$
  - Marks: $z_i \in \{$Lymphocytes (●), stroma cells (●), and epithelial cells (● and ●)$\}$

Data - Point Patterns w/ Shape

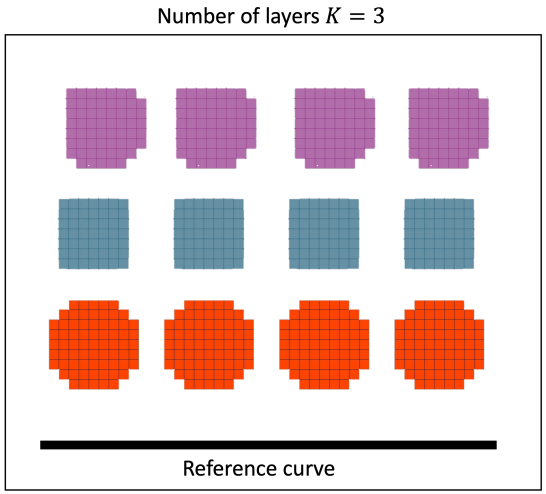
- Processed data: point pattern data w/ shape
  - ~~Locations of cell nuclei: $(x_i, y_i), i = 1, \ldots, n$~~
  - Marks: $z_i \in \{$~~Lymphocytes~~ (●), ~~stroma cells~~ (●), and epithelial cells (● and ●)$\}$
  - Cell nucleus pixel: $(x_{ij}, y_{ij}), i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, where m_i is the number of total pixels in cell nucleus i

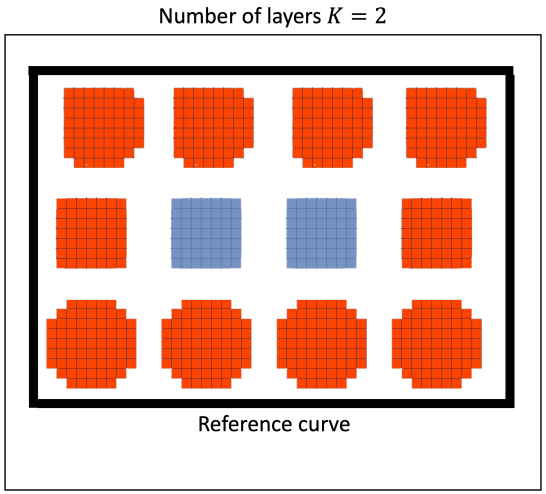


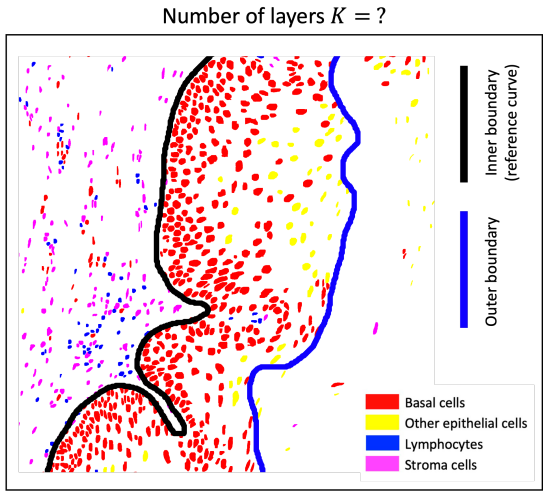


Data - Reference Curves

- Layered structure depends on the selection of reference curves
- Pathologists define the outer boundary near the corneum cells (layers are less distinct), and the inner boundary near the denser stromal regions

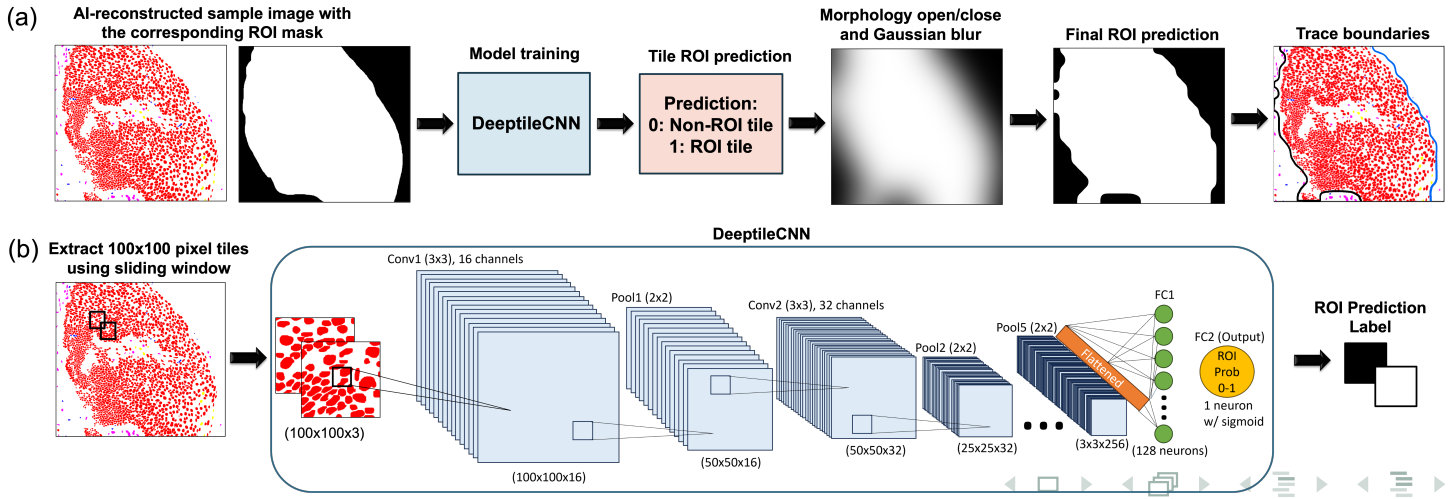






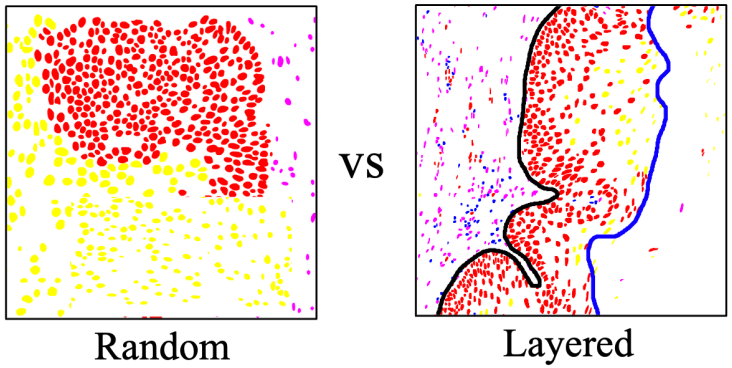
AI - DeeptileCNN

- Boundary detection is approached as tile classification
- Region of interest (ROI) classification using DeeptileCNN
  - Tiles were labeled as ROI if $\geq 75\%$ of pixels overlapped with the ground-truth ROI mask
  - Size: $55,613$ (training), $15,930$ (validation), and 7965 (test)
  - Overall accuracy: 87.1% (training) and 85.5% (test)
  - Gaussian smoothing was used to reduce blocky edges by tiling

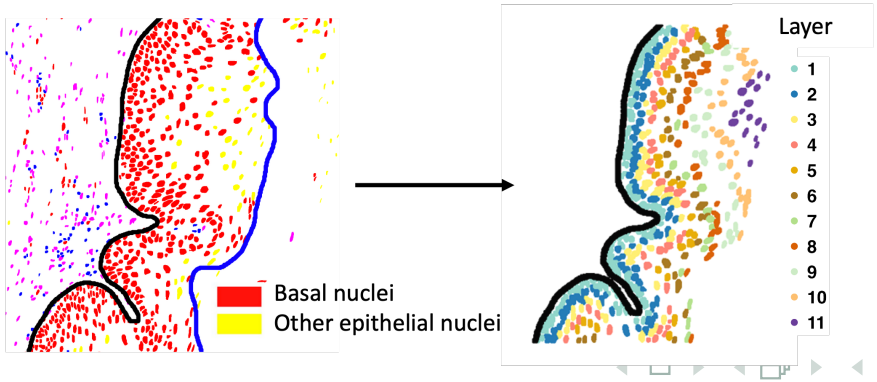


Research Goals

- To study the layered structure in point pattern data
  - To test if point pattern data exhibit a layered structure

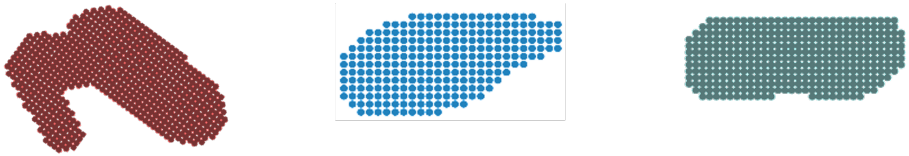


  - To estimate the number of layers from layered point pattern data

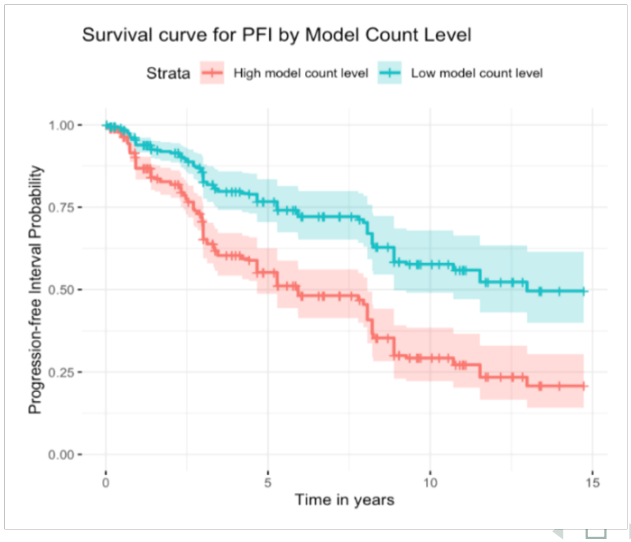


Research Goals

- To characterize shape of cell nuclei

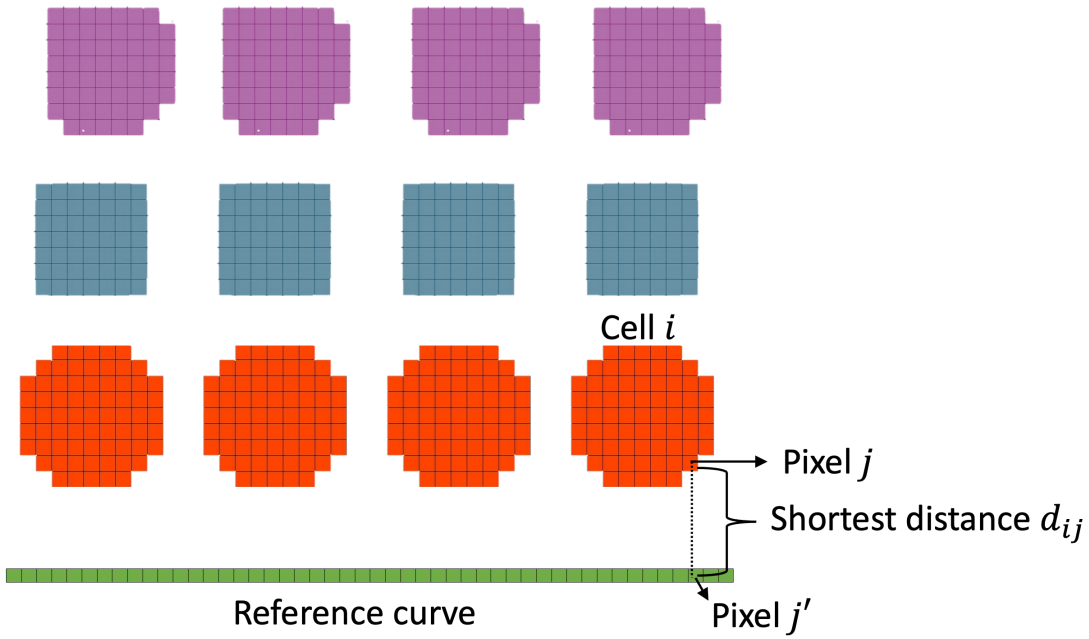


- To assess the clinical significance of epithelial layer numbers and oral cancer progression

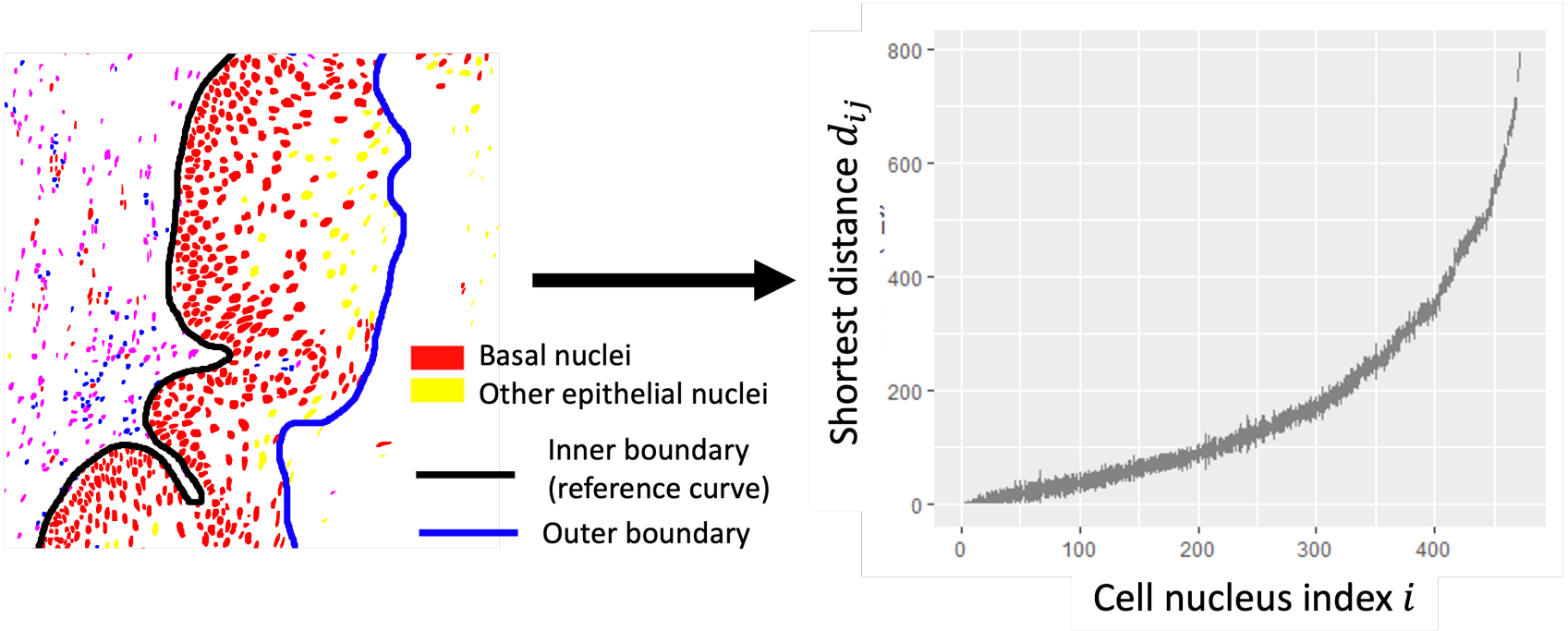


BLADE - Input Data

- Input data d_{ij}: the shortest Euclidean distance from the pixel j in cell nucleus i to the reference curve
- $\boldsymbol{d}_{i\cdot} = (d_{i1}, \ldots, d_{i,m_i})$ captures the spatial position of each cell nucleus i to the reference curve

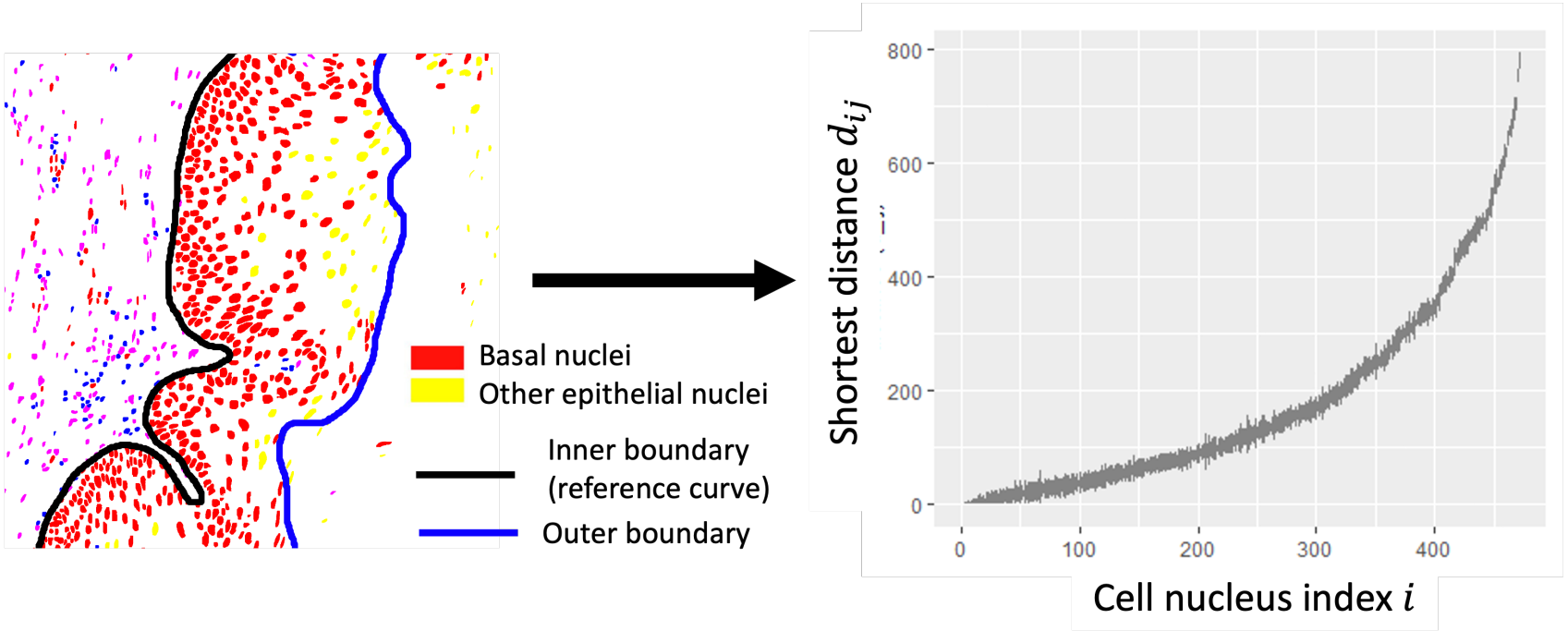


BLADE - Input Data



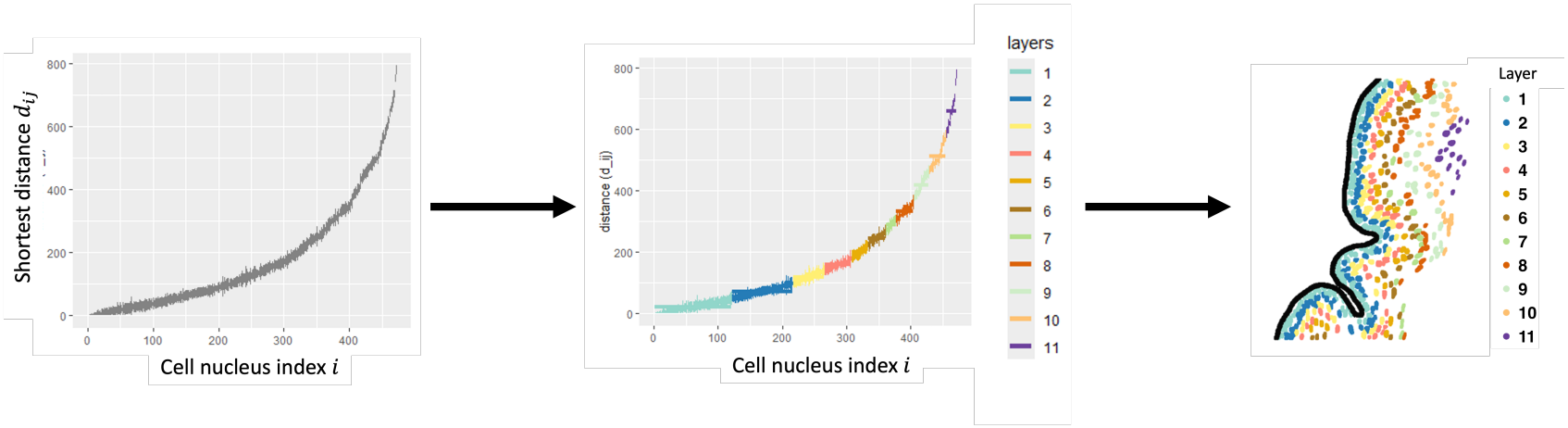
- Each bar represents a cell nucleus i
- The two ends of each bar i indicating the range of shortest distances $\boldsymbol{d}_{i\cdot} = (d_{i1}, \ldots, d_{im_i})$

BLADE - Input Data



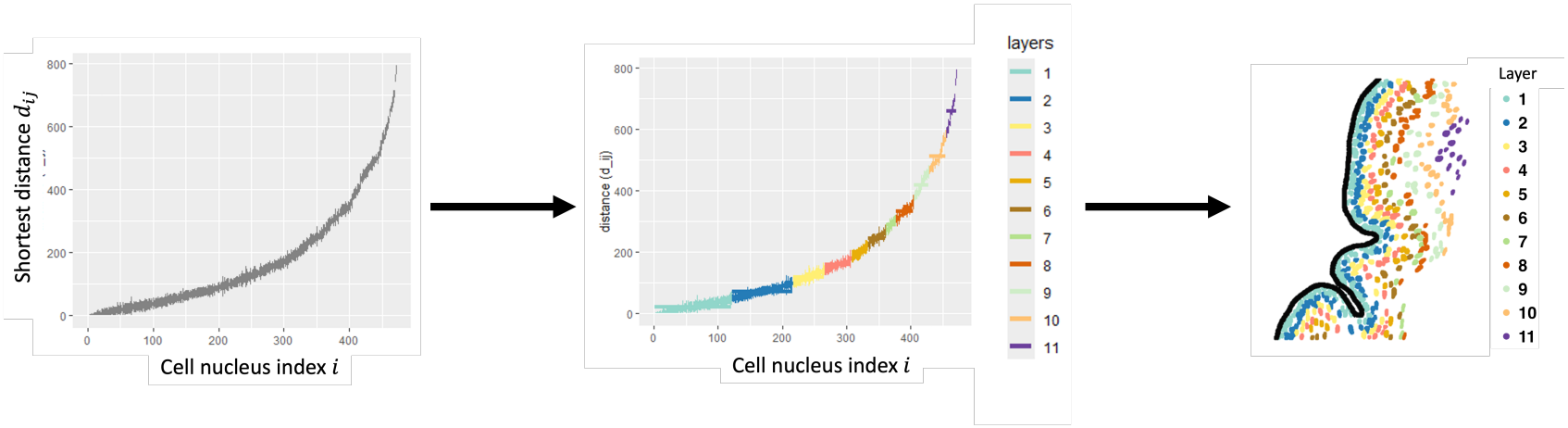
- Each bar represents a cell nucleus i
- The two ends of each bar i indicating the range of shortest distances $\boldsymbol{d}_{i\cdot} = (d_{i1}, \ldots, d_{im_i})$
- View $\boldsymbol{d}_{i\cdot}$ as sampling points from a function

BLADE - Overview



- Cluster the functions with sampling points $\boldsymbol{d}_{i\cdot}, i = 1, \ldots, n$ into distinct groups in a sequential order
  - Not necessary to include all elements in $\boldsymbol{d}_{i\cdot}$ to reconstruct the nucleus shape

BLADE - Overview



- Cluster the functions with sampling points $\boldsymbol{d}_{i\cdot}, i = 1, \ldots, n$ into distinct groups in a sequential order
  - Not necessary to include all elements in $\boldsymbol{d}_{i\cdot}$ to reconstruct the nucleus shape
- What is the most appropriate function?
- What is the most appropriate clustering approach?

BLADE - Shape Component

- Assume d_{ij} follows a generalized Beta (GBe) distribution:

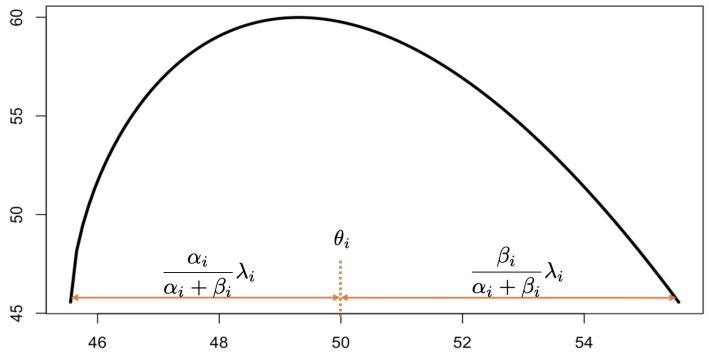
$$d_{ij}|\alpha_i, \beta_i, \theta_i, \lambda_i \sim \text{GBe}(\alpha_i, \beta_i, \theta_i, \lambda_i)$$

  - p.d.f.: $f(d_{ij}|\alpha_i, \beta_i, \theta_i, \lambda_i) =$

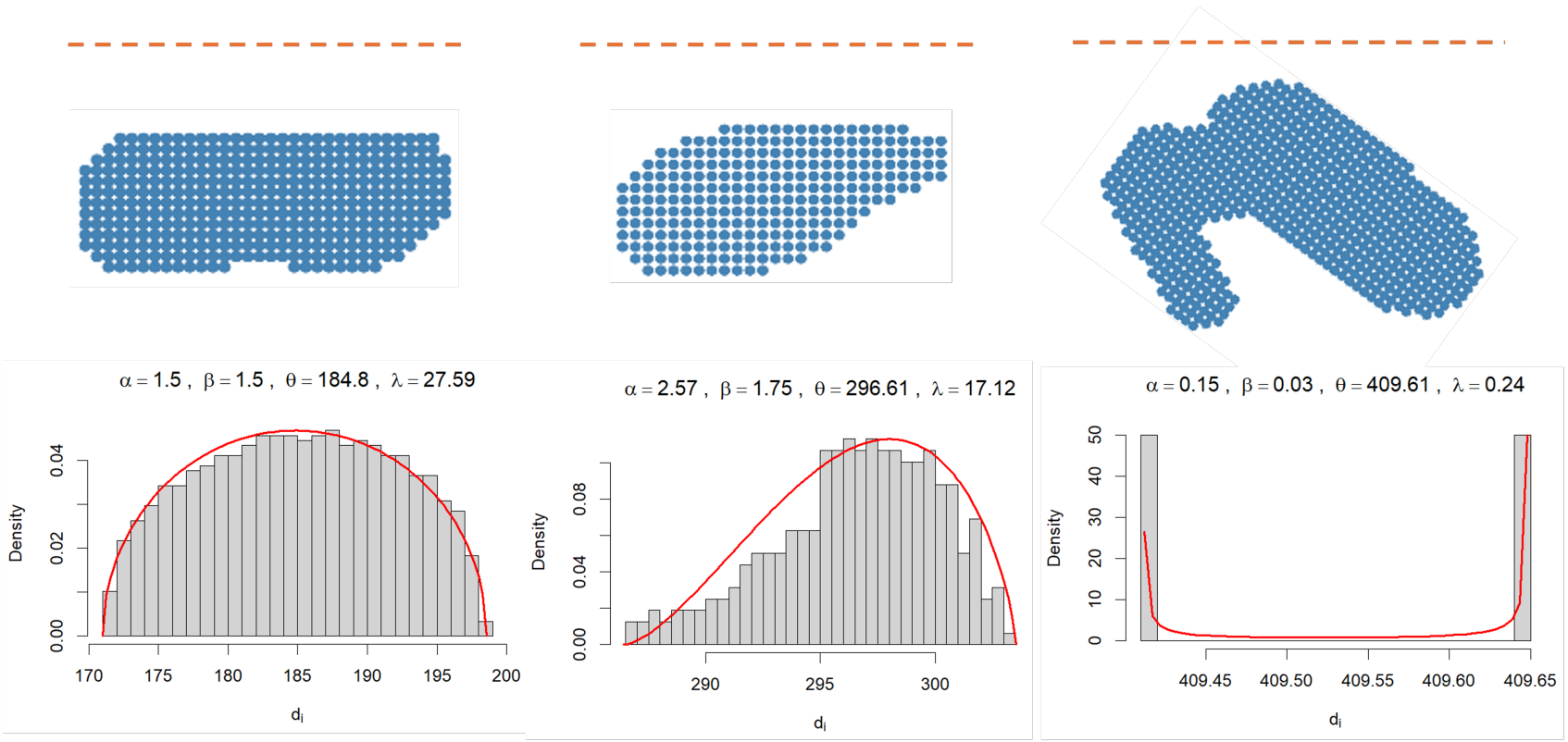
$$\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \frac{\left(d_{ij}-\theta_i+\frac{\alpha_i}{\alpha_i+\beta_i}\lambda_i\right)^{\alpha_i-1}\left(-d_{ij}+\theta_i+\frac{\alpha_i}{\alpha_i+\beta_i}\lambda_i\right)^{\beta_i-1}}{\lambda_i^{\alpha_i+\beta_i-2}}$$

  - Relationship with the Beta distribution:

$$\frac{d_{ij}-\left(\theta_i-\frac{\alpha_i}{\alpha_i+\beta_i}\lambda_i\right)}{\lambda_i} \sim \text{Beta}(\alpha_i, \beta_i)$$

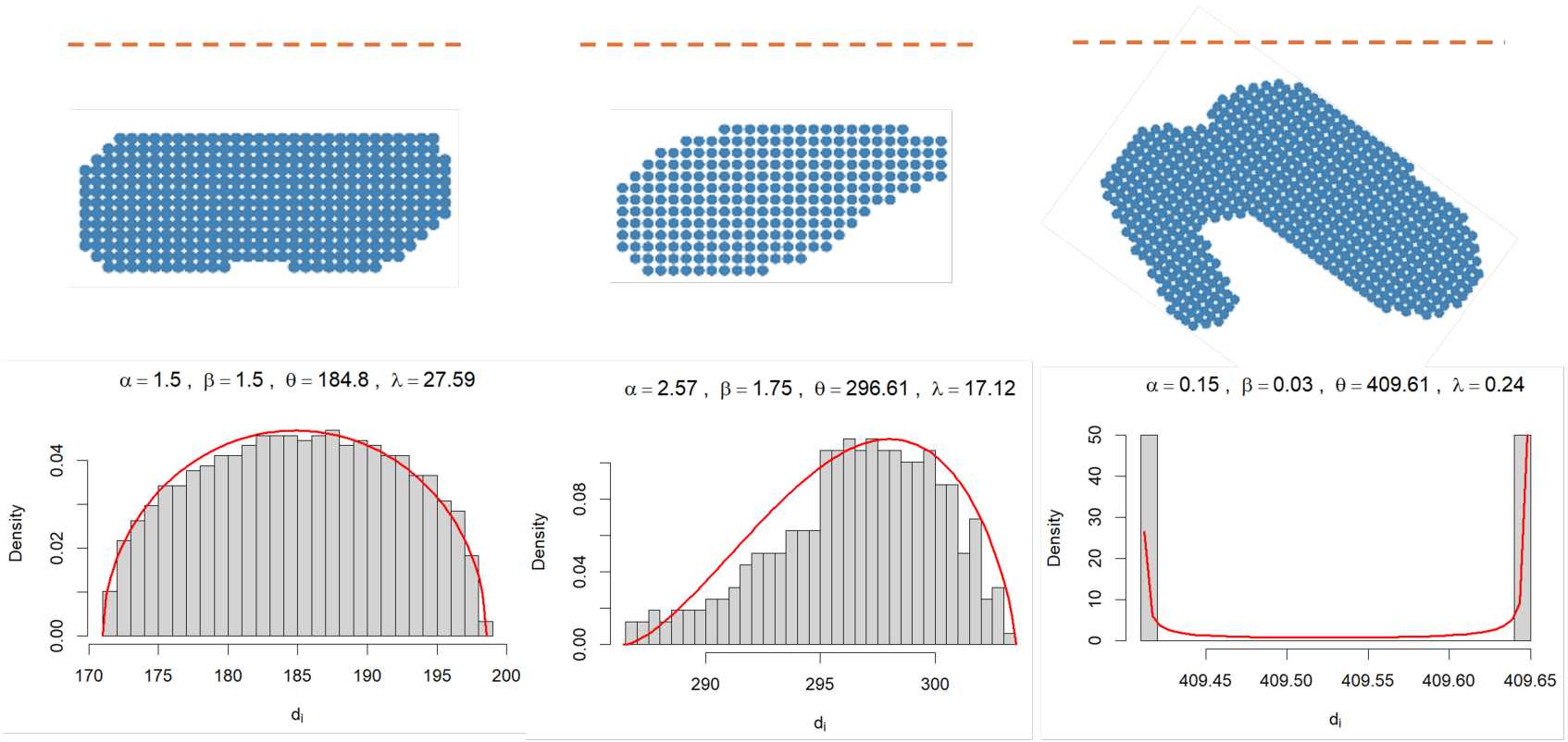


BLADE - Shape Component



- The estimated shape parameters α_i and β_i for three representative cell nuclei

BLADE - Shape Component



- The estimated shape parameters α_i and β_i for three representative cell nuclei
- Extended to α_k and β_k, assuming nuclei shape is similar

BLADE - FMM

- Adjust for the shape effect *via* the GBe model
- The stratification focuses on clustering the mean shortest distances θ_i's

BLADE - FMM

- Adjust for the shape effect *via* the GBe model
- The stratification focuses on clustering the mean shortest distances θ_i's
- Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ denote the latent layers, where $z_i = k$ indicates the cell nucleus i belongs to layer k
- GMM with conjugate prior

$$\begin{aligned} \theta_i | z_i = k, \mu_k, \sigma_k^2 &\sim \mathsf{N}(\mu_k, \sigma_k^2) \\ \mu_k, \sigma_k^2 &\sim \mathsf{NIG}(\mu_0, p_0, \nu_0/2, SS_0/2) \\ z_i | \boldsymbol{\omega} &\sim \mathsf{Multi}(1, \boldsymbol{\omega}), \boldsymbol{\omega} \sim \mathsf{Dir}(\boldsymbol{\gamma}) \end{aligned}$$

BLADE - DPMM

- In the Dirichlet process (DP), the weight ω_k is constructed *via* stick-breaking process:

$$\omega_k = v_k \prod_{h=1}^{k-1}(1 - v_h), \quad v_k \sim \text{Be}(1, \gamma)$$

  - γ is the concentration parameter that controls the variation of the DP prior around its mean G_0

- DPMM

$$\theta_i | \mu_i, \sigma_i^2 \sim \text{N}(\mu_i, \sigma_i^2)$$
$$\mu_i, \sigma_i^2 | G \sim G$$
$$G \sim \text{DP}(G_0, \gamma), \text{where } G_0 = \text{NIG}(\mu_0, p_0, \nu_0/2, SS_0/2)$$

BLADE - DPMM

- The random partition can be generated with an urn scheme or a Chinese restaurant process (CRP)
  - The conditional distribution for each z_i is:

$$\Pr(z_i = k \mid \boldsymbol{z}_{-i}) \propto \begin{cases} n_{k,-i} & \text{for an existing cluster} \\ \gamma & \text{if } c \text{ is a new cluster} \end{cases},$$

  - As the sample size grows, $\hat{K}$ remains inconsistent due to the presence of extraneous clusters in the posterior
  - Flexible but often leads to overestimation of clusters.

BLADE - MFM

- Mixture of Finite Mixtures (MFM) proposed by Miller and Harrison (*JASA*, 2018)

$$z_i|K,\boldsymbol{\pi} \sim \sum_{k=1}^{K} \pi_k I(z_i = k), \quad \boldsymbol{\pi}|K \sim \text{Dir}(\boldsymbol{\gamma}), \quad K-1 \sim \text{Poi}(\tau)$$

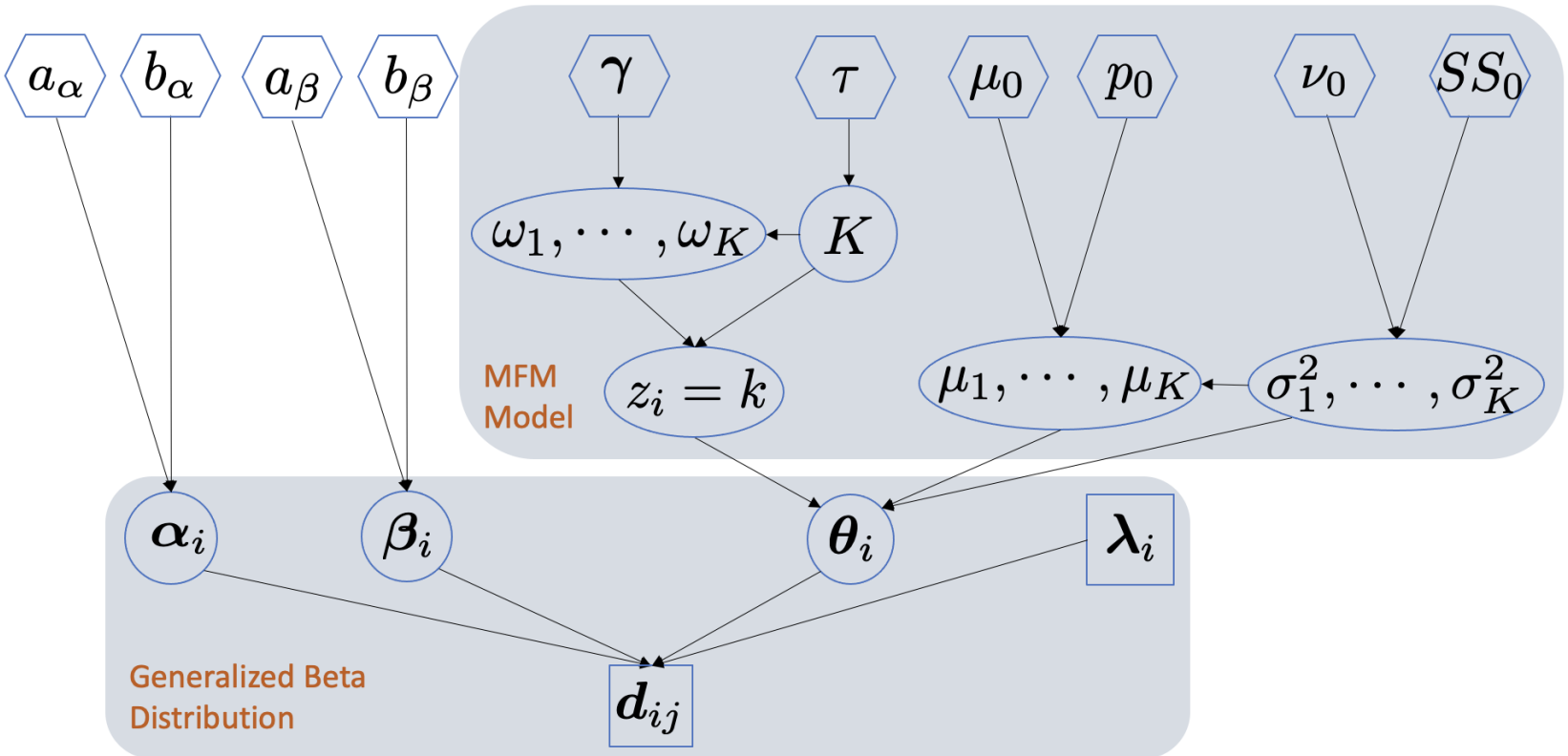
  - The conditional distribution for each z_i under MFM is:

$$\Pr(z_i = k \mid \boldsymbol{z}_{-i}) \propto \begin{cases} n_{k,-i} + \gamma & \text{for an existing cluster } k \\ \frac{V_n(|\boldsymbol{z}_{-i}|+1)}{V_n(|\boldsymbol{z}_{-i}|)}\gamma & \text{if } k \text{ is a new cluster} \end{cases}$$

  - Reduces the likelihood of forming extraneous clusters as $\frac{V_n(|\cdot|+1)}{V_n(|\cdot|)} < 1$
  - Avoids overestimation as number of clusters converges to a finite value

BLADE - Wrap-up

- Graphical representation:

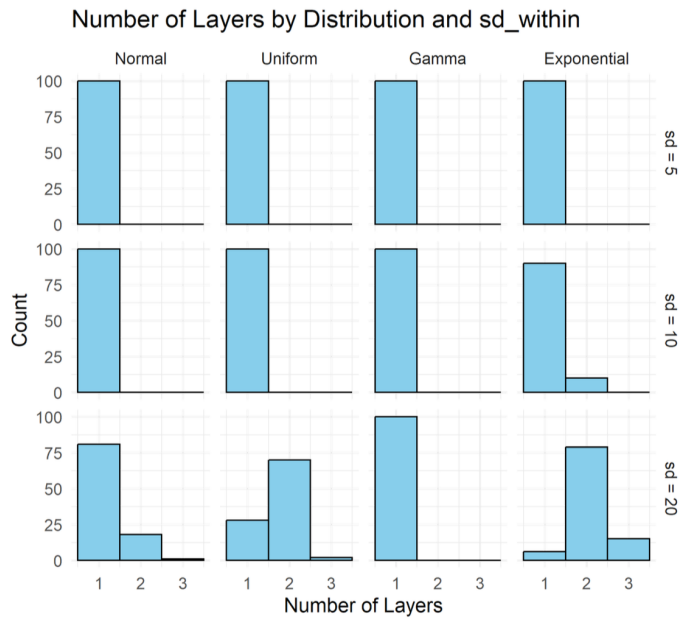


BLADE - Hypothesis Testing

- To test whether a given point pattern exhibits a layered structure or not *via* Bayes factor (BF)

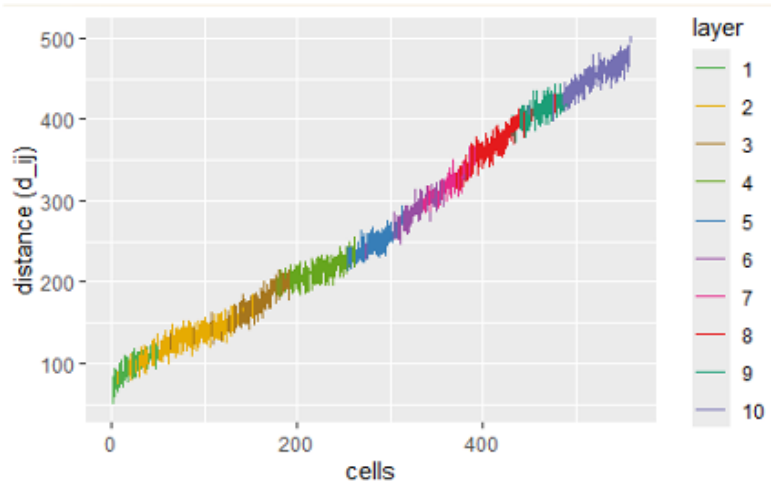
$$\text{BF} = \frac{\Pr(K=1|\cdot)/\Pr(K\geq 2|\cdot)}{\Pr(K=1)/\Pr(K\geq 2)} = \frac{\sum_{t=1}^{T} I(|\boldsymbol{z}^{(t)}|=1)/I(|\boldsymbol{z}^{(t)}|\geq 2)}{e^{-\tau}/(1-e^{-\tau})}$$

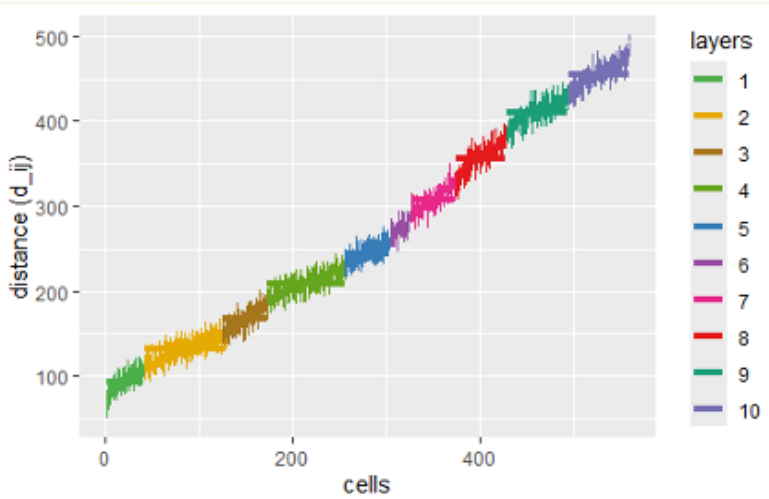
  - When MFM prior assumes $K-1 \sim \text{Poi}(\tau)$



BLADE - Simulation

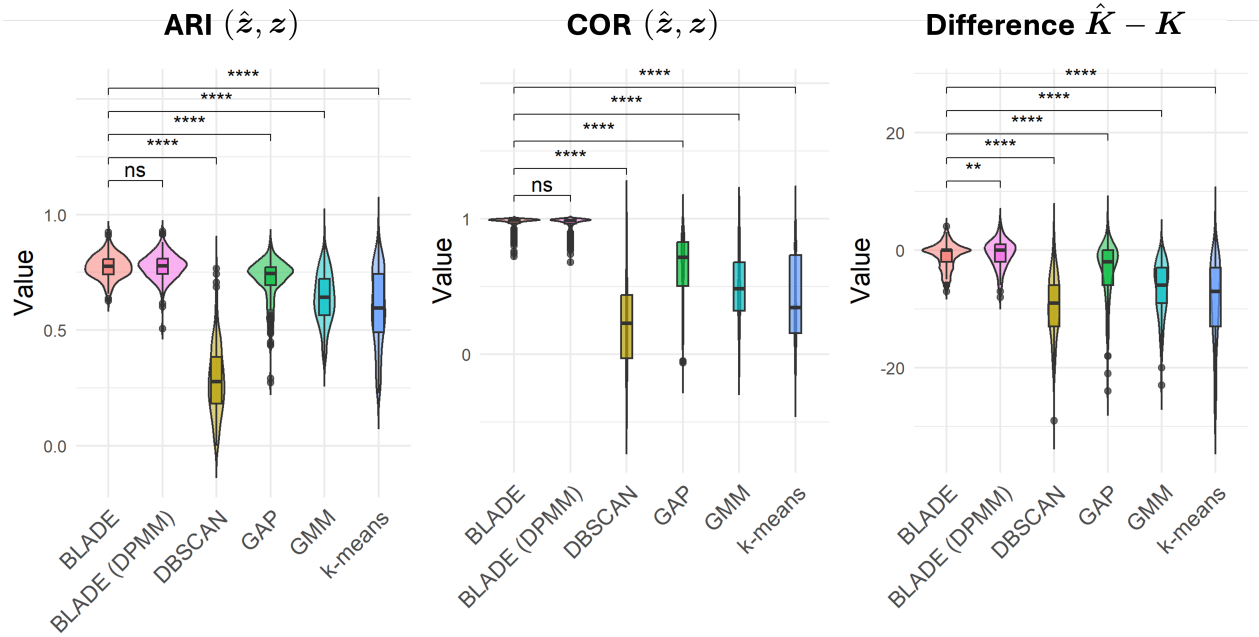
- Simulation setup:
  - Generated 200 simulated datasets
  - $K_{\text{true}} \sim \text{U}(5, 35)$, $n_k \sim \text{N}_{[20,200]}(100, 60^2)$, and $m_i \sim \text{U}(90, 150)$





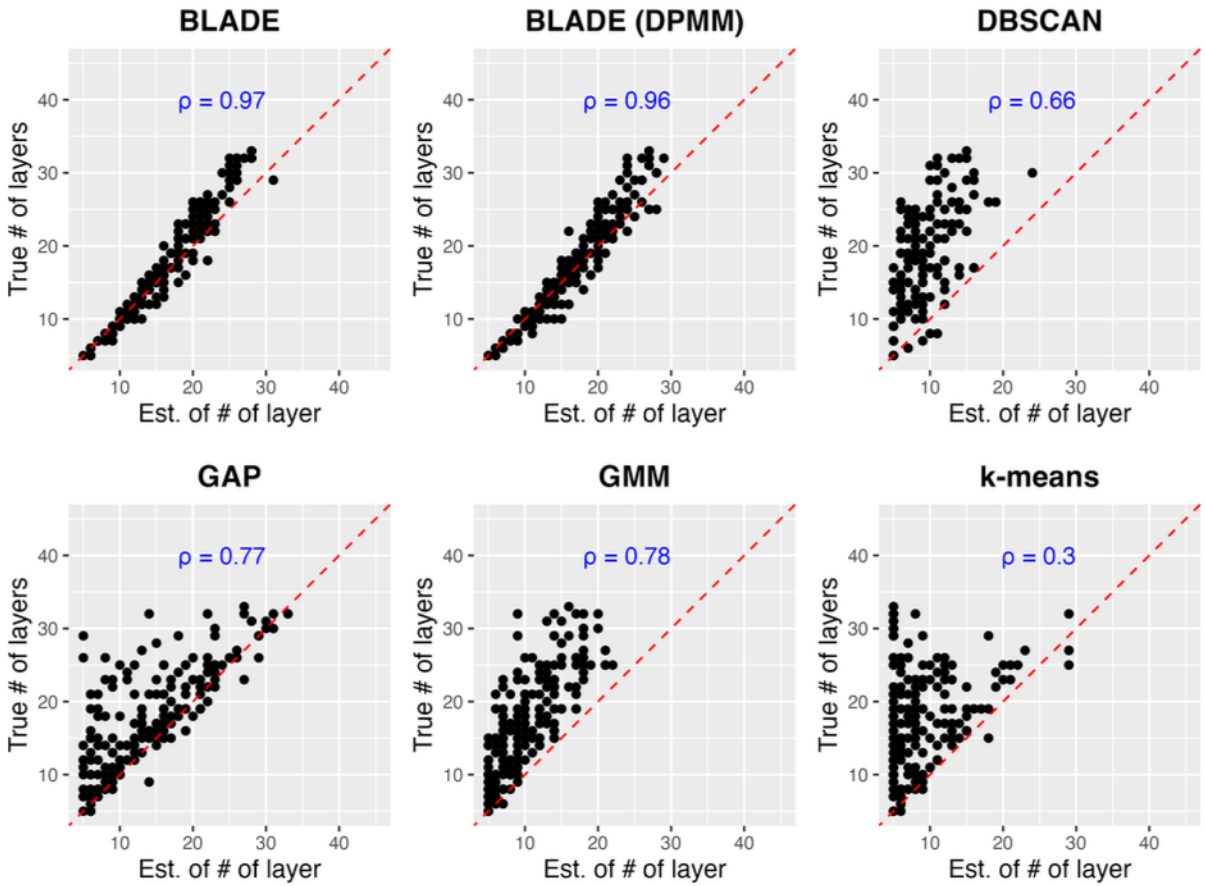
BLADE - Simulation

- Comparison of layered structures estimated by BLADE and other methods



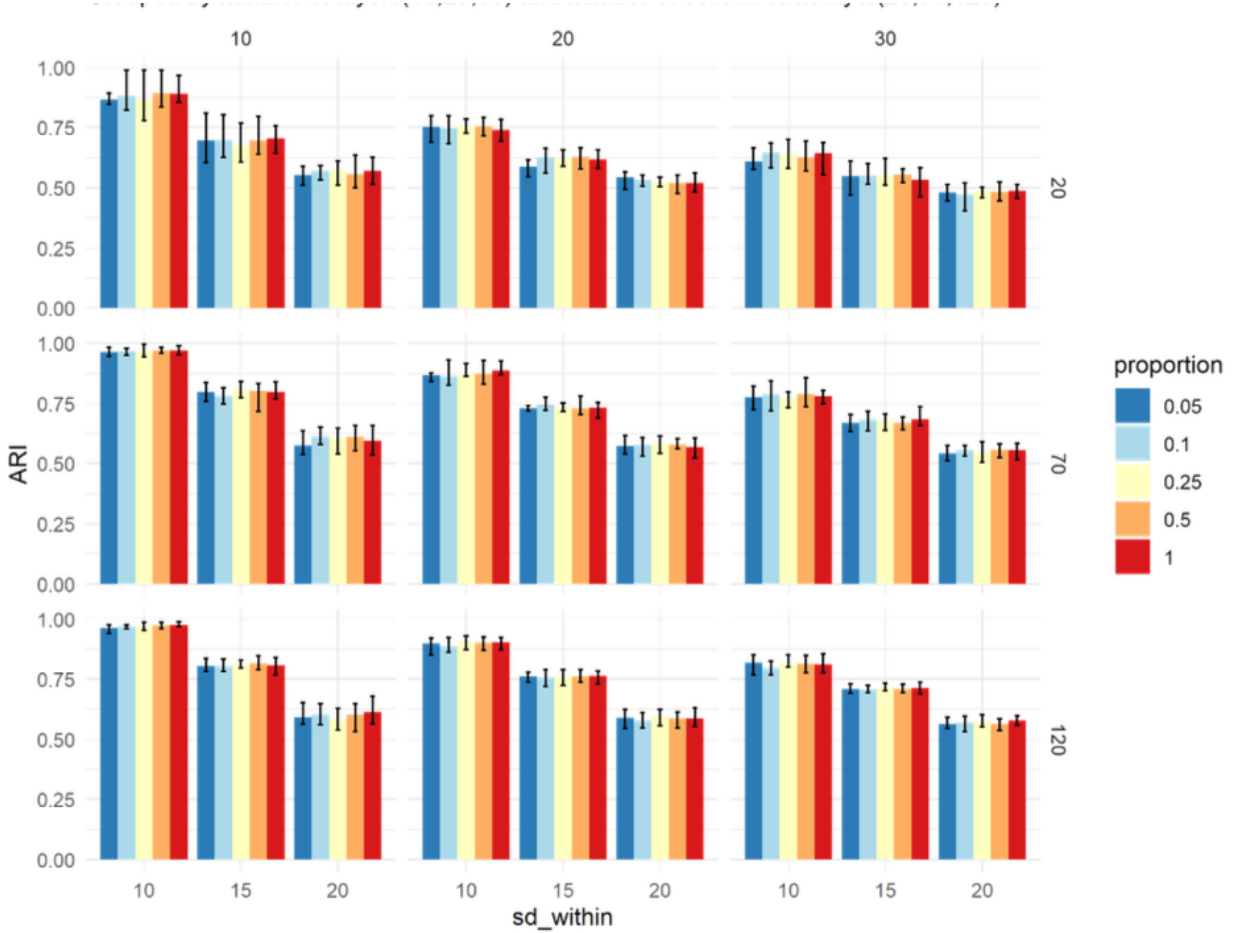
BLADE - Simulation

- Comparison of layered structures estimated by BLADE and other methods



BLADE - Simulation

- BLADE maintained high ARI scores even when only a small proportion of pixel-level data d_{ij} were used.



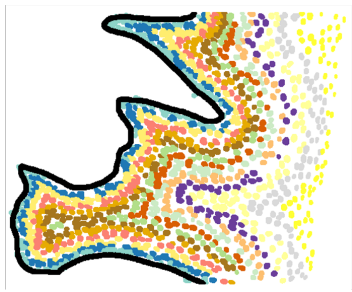
BLADE - Oral Tissue Pathology Images

- Oral cancer is the sixth most common cancer globally
- Early detection increases survival: 83.7% (early) vs. 38.5% (late stage)
- 128 patients with oral potentially malignant disorders (OPMD) from the erlotinib prevention of oral cancer (EPOC) trial at UT MDACC
- Clinical Data:
  - Demographics: age, gender
  - Risk factors: prior/concurrent oral cancer, TP53 mutation, LOH eligibility, leukoplakia group, HistBaselineR3
  - Outcomes: progression-free interval (PFI) status and time: $1 =$ progression observed and 0 otherwise

BLADE - Oral Tissue Pathology Images

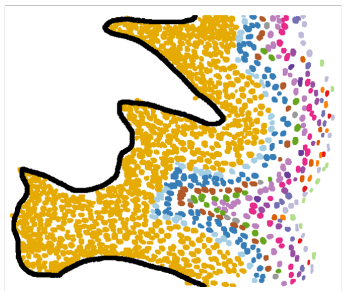
- The layered structures estimated by BLADE and the other methods for patient 076

BLADE



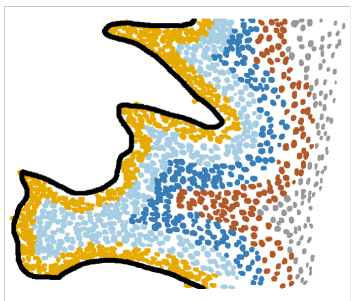
$\hat{K} = 14$

DBSCAN



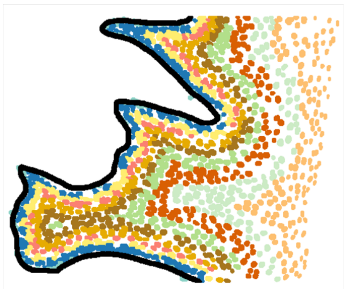
$\hat{K} = 16$

GAP



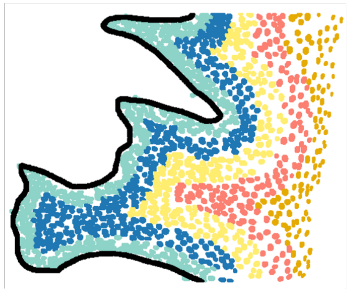
$\hat{K} = 5$

GMM



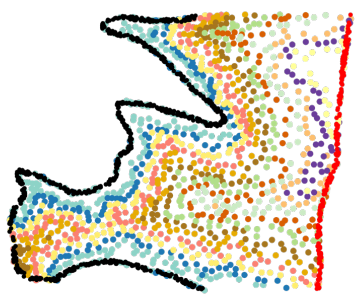
$\hat{K} = 10$

k-means



$\hat{K} = 5$

OPM



$\hat{K} = 12$

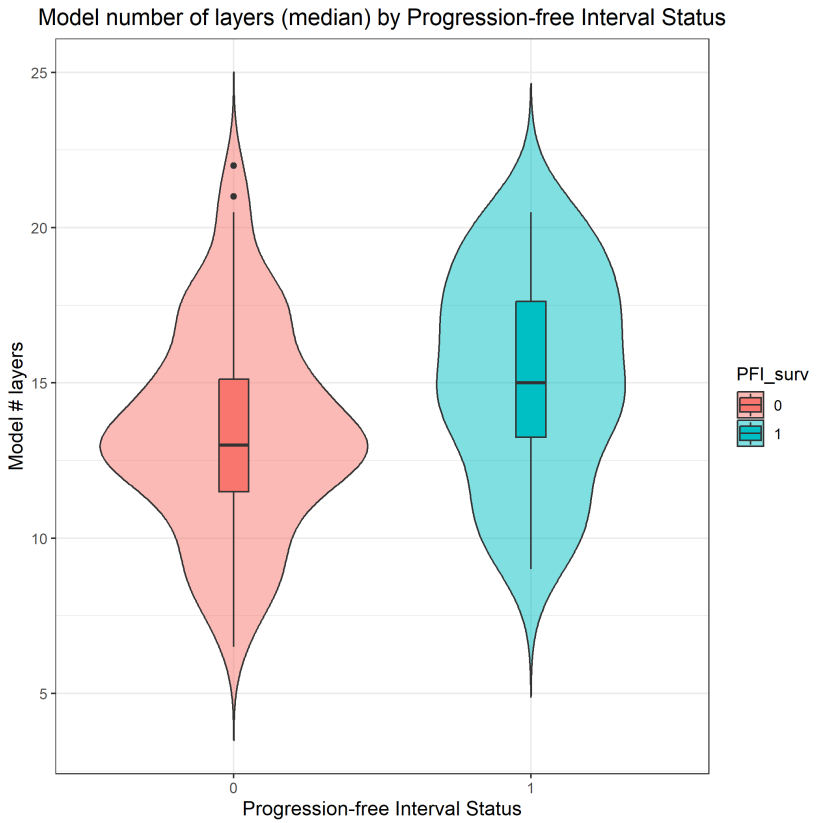
BLADE - Oral Tissue Pathology Images

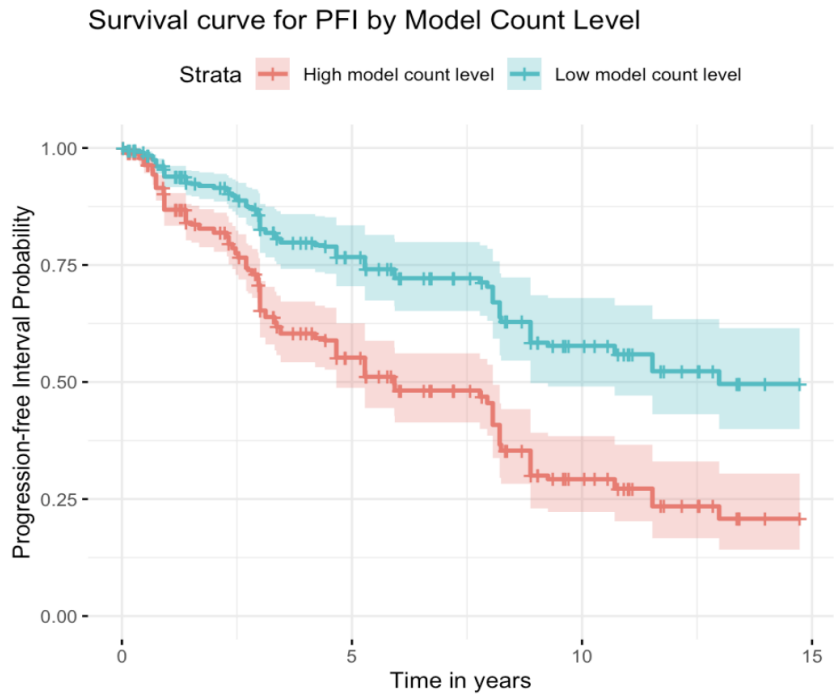
- Model-derived layer count is significantly associated with disease progression
- Each additional layer increases the risk of progression by approximately 8%

Variable	**Coefficient**	**Exp(Coef.)**	**SE**	**Pr($>|z|$)**
$\hat{K}$ by BLADE	0.0771	**1.0801**	0.0237	**0.00113**
Age	-0.0302	0.9703	0.0156	0.05370
Male vs. female	-0.4778	0.6201	0.4149	0.24945
Prior oral cancer	0.9873	2.6841	0.3854	**0.01042**
Concurrent oral cancer	0.7515	2.1203	0.6273	0.23091
Leukoplakia group	0.4116	1.5092	0.6335	0.51589
LOH Eligible	0.4665	1.5944	0.3864	0.22727

BLADE - Oral Tissue Pathology Images

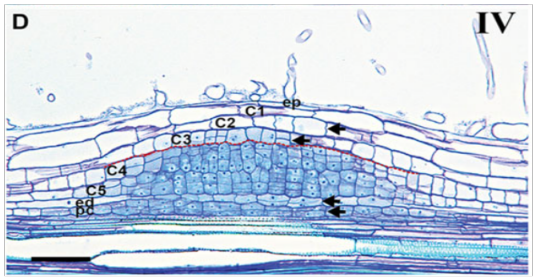
- Patients with PFI status $= 1$ exhibited higher median layer counts compared to those with PFI status $= 0$

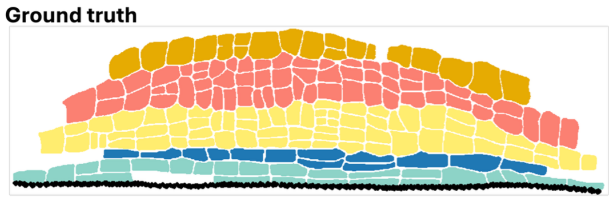


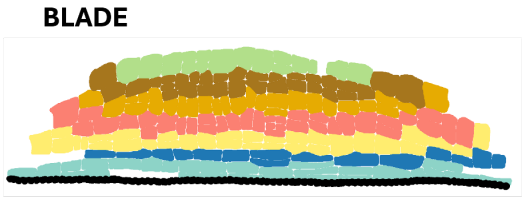


BLADE - *Medicago Truncatula* Root Nodule

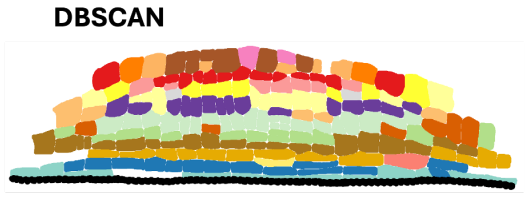
- The layered structures at STAGE IV estimated by BLADE and the other methods



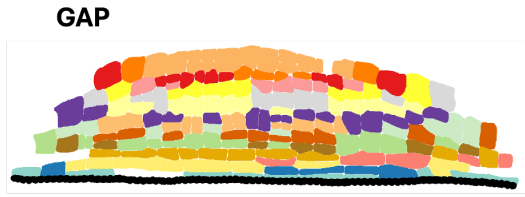




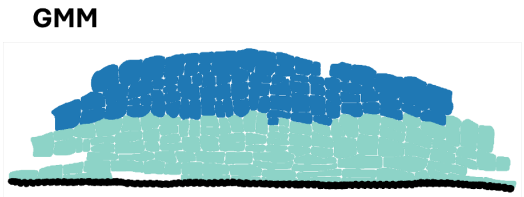
$\hat{K} = 7, \quad \text{ARI} = 0.489, \quad \text{COR} = 0.929$



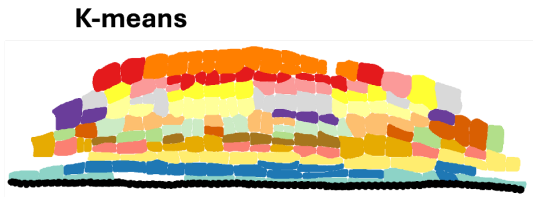
$\hat{K} = 20, \quad \text{ARI} = 0.272, \quad \text{COR} = 0.870$



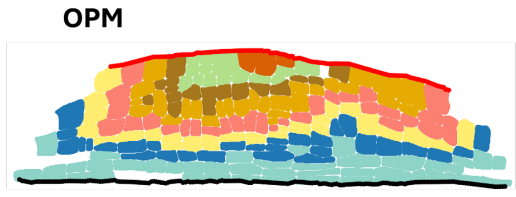
$\hat{K} = 18, \quad \text{ARI} = 0.204, \quad \text{COR} = 0.263$



$\hat{K} = 2, \quad \text{ARI} = 0.425, \quad \text{COR} = 0.087$



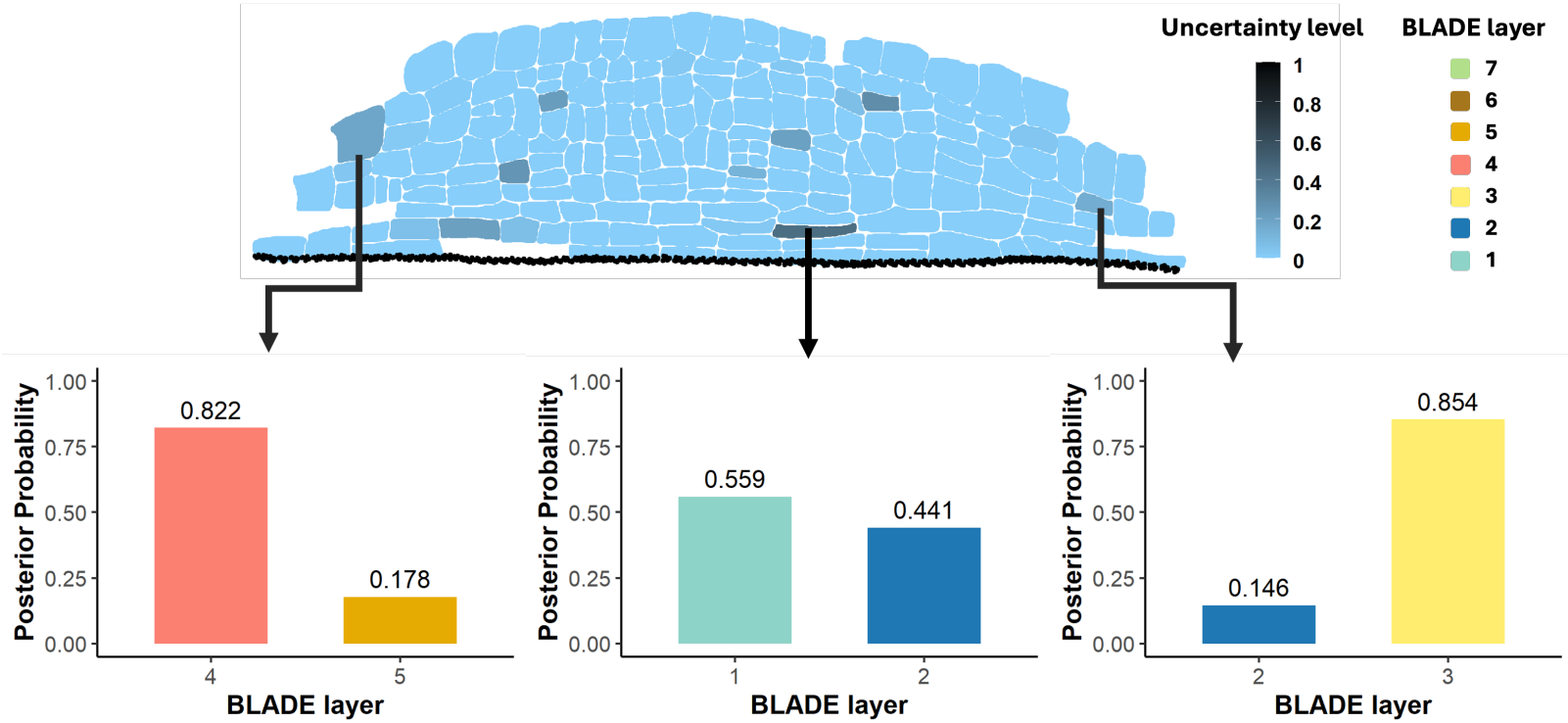
$\hat{K} = 17, \quad \text{ARI} = 0.215, \quad \text{COR} = 0.523$



$\hat{K} = 8, \quad \text{ARI} = 0.212, \quad \text{COR} = 0.823$

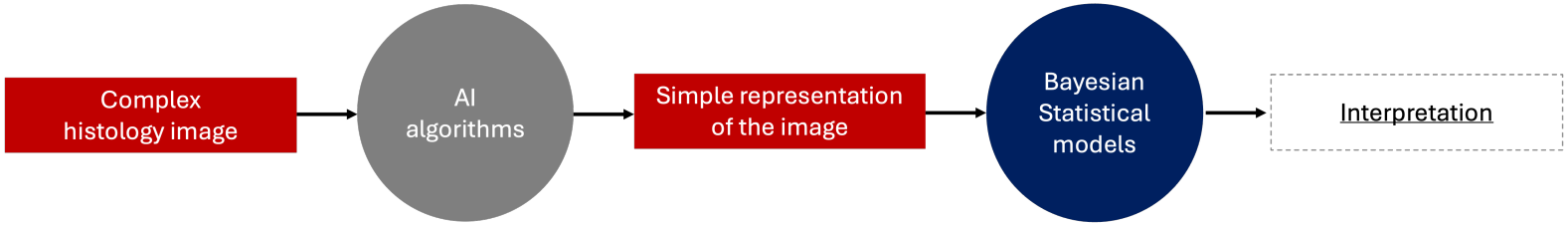
BLADE - *Medicago Truncatula* Root Nodule

- BLADE can also quantify the uncertainty, which can be defined as the posterior probability $\Pr(z_i \neq \hat{z}_i|\cdot)$



Summary

- Statistical shape and spatial analysis under the AI-Statistics workflow



- Bayesian nonparametric model for analyzed layered point pattern data
  - Test if point pattern data exhibit a layered structure
  - Estimate the number of layers from point pattern data
  - Effectively characterize cell nucleus shape information

The End

Thanks for your attention!

- Acknowledgment

DMS-2210912
DMS-2113674

1R01GM141519
1R01DK131267