Statistical Inference for Spatial Transcriptomics in the Age of Deep Learning

Roman Kouznetsov

University of Michigan

September 30, 2024

Outline

1. Background
2. Our Model (DeepST)
3. Results
4. Conclusion

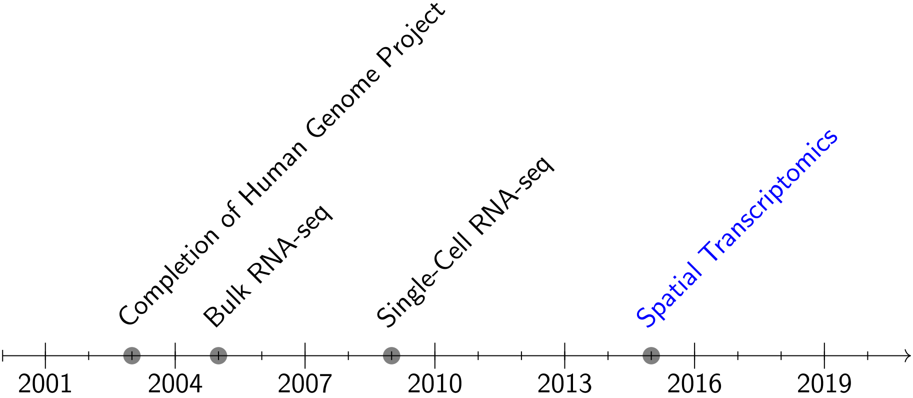
1 Background

2 Our Model (DeepST)

3 Results

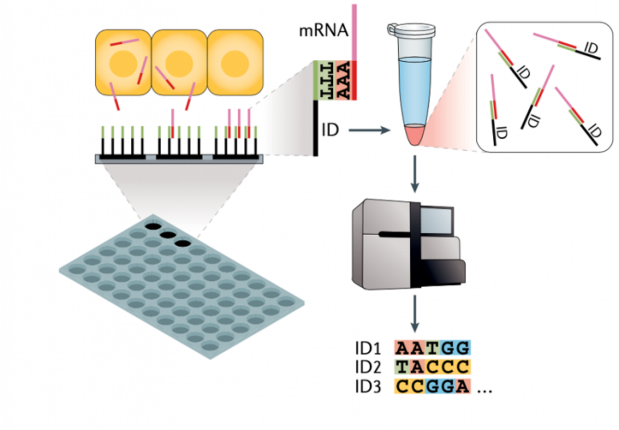
4 Conclusion

History of Gene Sequencing



Spatial Transcriptomics

- Spatial Transcriptomics (ST) ties cell expressions to cell positions.
- Prior to ST one could not get single-cell resolution of position and expression pairings.



Cell-Cell Communication (CCC)

- Cells communicate with one another, creating gene pathways.
- Cells send signals (**ligands**), and their neighbors collect those signals using receivers (**receptors**).
- When a cell receives a signal, its own expressions can change. **We would like to model this behavior.**

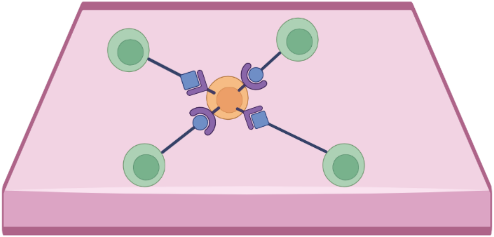
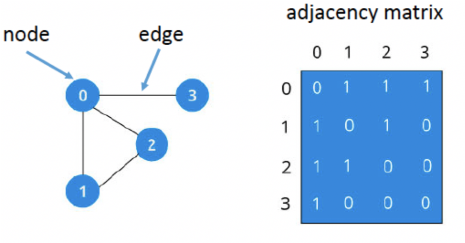


Figure: **Target Cell** (orange) receiving signals from **signalling cells** (green).

Graphs

- Graph: (Nodes, Edges)
- G = (V, E)
- Edges can be expressed in a matrix called an adjacency matrix (A).
- Each node can have attributes that contain pertinent information about a specific node.



adjacency matrix

	0	1	2	3
0	0	1	1	1
1	1	0	1	0
2	1	1	0	0
3	1	0	0	0

Image Source: Andy Jahn

Graph Convolutional Networks (GCNs)

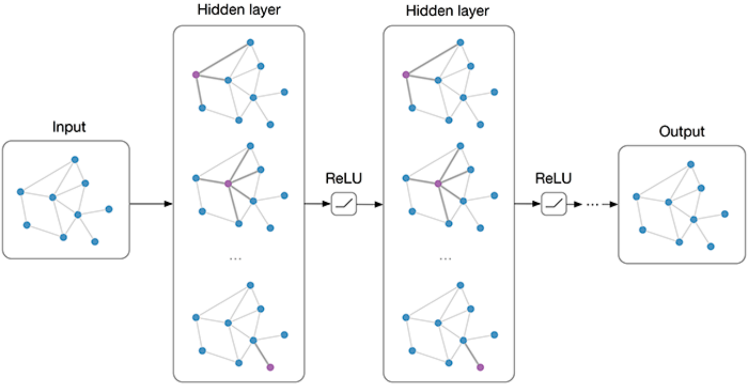


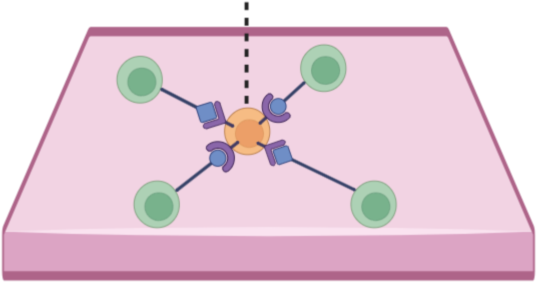
Image Source: Thomas Kipf

Gaussian Mixture Model Convolutions (GMMConv)

- A convolution that treats each neighboring signal as a mode in a GMM.
- K: number of Gaussian kernels
- Θ_k: the weights of a dense graph neural network
- $e_{i,j}$: pseudo-coordinates for the pair (cell i, cell j)
- w_k: weighting function (kernel)
- $\mathcal{N}(i)$: the neighbors of target cell i

$$\mathbf{x}_i' = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \frac{1}{K} \sum_{k=1}^{K} \mathbf{w}_k(\mathbf{e}_{i,j}) \odot \mathbf{\Theta}_k \mathbf{x}_j$$

$$\mathbf{w}_k(\mathbf{e}) = \exp\left(-\frac{1}{2}(\mathbf{e} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{e} - \mu_k)\right)$$



Orange: **Target** Cell, Green: **Neighboring** Cells

1 Background

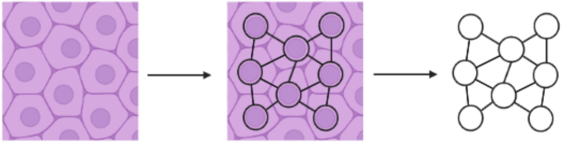
2 Our Model (DeepST)

3 Results

4 Conclusion

Tissues as Graphs

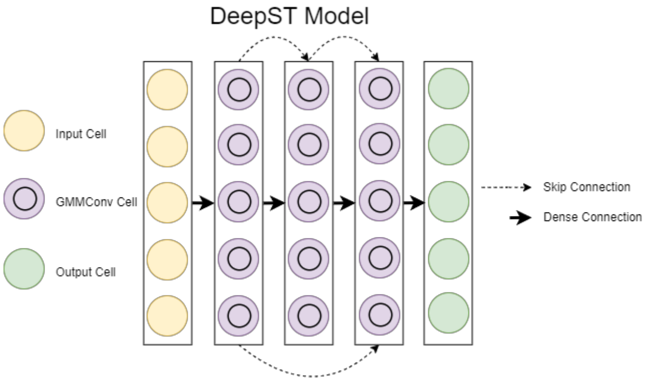
- Tissue samples can be represented as graphs!
- Let cells represent nodes.
- Let cell-cell communications represent edges.



- Given cells i and j, we consider these cells to have the following CCC structure:

$$A_{ij} = A_{ji} = \begin{cases} 1 & d(i,j) \leq r \\ 0 & d(i,j) > r \end{cases}$$

DeepST



One-Dimensional Schematic of the DeepST Model

1 Background

2 Our Model (DeepST)

3 Results

4 Conclusion

Application on MERFISH Hypothalamus Data

- 181 tissues of ST Data
- 36 animals
- $\approx$ 1 million cells
- 161 genes
  - 31 receptors (R)
  - 40 ligands (L)
  - 84 responses (genes that are neither ligands nor receptors) (Y)
  - 6 blanks
- $\mathcal{N}(L)$, and $\mathcal{N}(R)$ are the neighboring ligand and receptor expression respectively as defined by the graph.
- **Goal**: Model the regression $Y \sim \mathrm{DeepST}(L, R, \mathcal{N}(L), \mathcal{N}(R))$

MERFISH Results: Improved Prediction

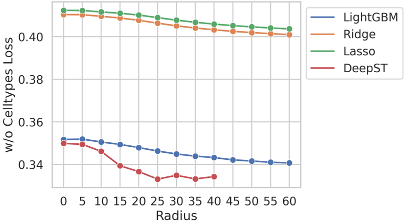


Figure: Without Cell Types

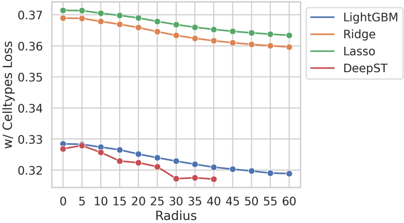
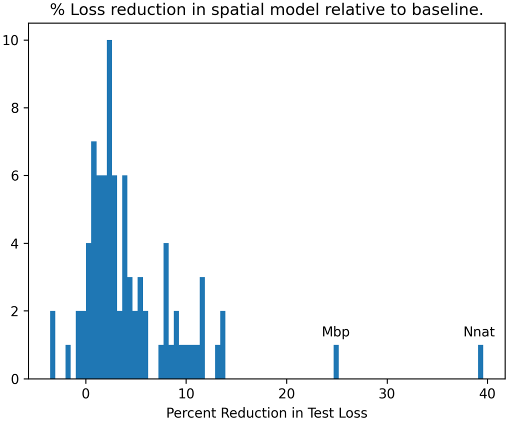


Figure: With Cell Types

Figure: MSE for models without cell types in input (left) and with cell types in input (right).

MERFISH Results: Spatially Dependent Genes

- Looking at each response gene individually, we can see which genes are more accurately predicted with a spatial model and by how much.



Semi-Synthetic Experiments

- Simulated expressions with real positions collected from ST data.
- Allows us to evaluate a wide array of expression circumstances to stress test model performance.
- For the notation going forward, we represent X_{cg} to be the expression of gene g in cell c.
- For the semi-synthetic experiments that follow we simulate all gene expressions in the dataset $X_{cg}\,\forall c, g$.
- In our cases, X_{c0} is given a special relationship with the other expressions and is the only response gene we model for simplicity.
- In all of our experiments, we simulate the data with a **true radius of 30μm**.

Synthetic Experiment 1

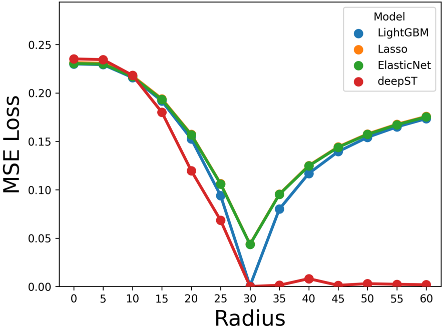


Figure: Synthetic Experiment #1 Test Losses

- $X_{cg} \sim \mathrm{NB}(1, 0.5)/5$
- $X_{c0} = \mathbb{1}\left(\sum_{c' \in \mathcal{N}(X_c)} X_{c'1} > 1\right) * \sum_{c' \in \mathcal{N}(X_c)} X_{c'1}$

Synthetic Experiment 2

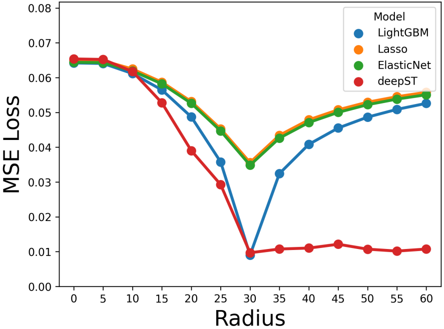


Figure: Synthetic Experiment #2 Test Losses

- $X_{cg} \sim \mathrm{Exp}(10)$
- $X_{c0} = X_{c0} + \mathbb{1}\left(\sum_{c' \in \mathcal{N}(X_c)} X_{c'1} > 1\right) * \sum_{c' \in \mathcal{N}(X_c)} X_{c'1}$

Synthetic Experiment 3

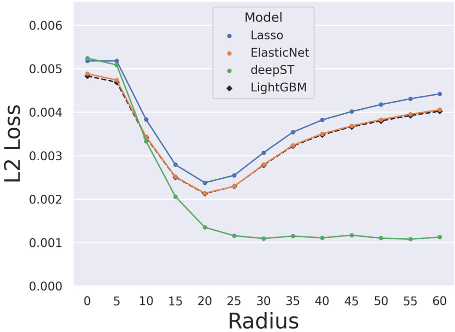


Figure: Synthetic Experiment #3 Test Losses

- $\mu_1, ..., \mu_G \sim N(20, 4)$
- $X_{cg} \sim NB(\mu_g, 0.5)/60$
- $X_{c0} = \sum_{X_{c'} \in \mathcal{N}(X_c)} \sqrt{X_{c'1}} \left(1 - \frac{\sinh^{-1}(5.863 * d(X_c, X_{c'}))}{5.863}\right)$

Causal Discovery with Measurement Error

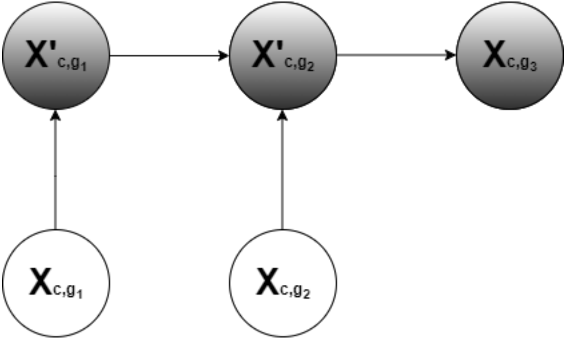


Figure: A probabilistic graph that shows why measurement error can prevent causal inference. White nodes represent the true values of covariates while grey nodes indicate noisy covariates resulting from measurement error.

False Discovery Synthetic Experiment

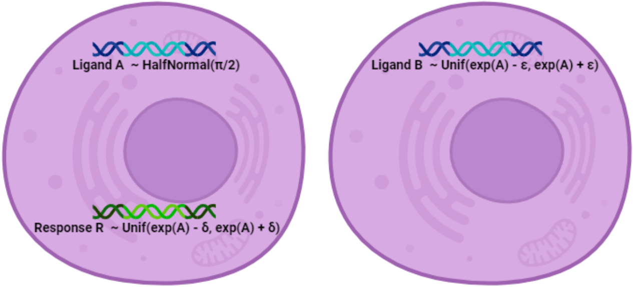
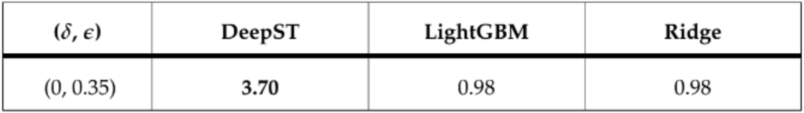

Ligand A ~ HalfNormal(π/2)
Response R ~ Unif(exp(A) - δ, exp(A) + δ)
Ligand B ~ Unif(exp(A) - ε, exp(A) + ε)

Figure: Target cell (left), Signalling cell (right)

- In the above simplified example, $R \perp B | A$.
- Therefore, an ideal model would inform us that spatial information is not relevant for inferring R.
- This can help us avoid spurious conclusions about spatial dependence.

False Discovery Synthetic Experiment (Results)



(δ, ϵ)	**DeepST**	**LightGBM**	**Ridge**
(0, 0.35)	**3.70**	0.98	0.98

Figure: Ratio of spatially aware loss to spatially ignorant loss across models. Best result for each setting is marked in bold.

Leveraging GVAEs

$$p(\mathbf{z}) = \mathcal{N}(0, I) \longrightarrow p(z_c) = N(0, I)$$

$$(\mu, \sigma) = \text{Encoder}_\phi(\mathbf{x}) \longrightarrow (\mu_c, \sigma_c) = \mathrm{DeepST}(X_{c,L}, X_{c,R}, X_{\mathcal{N}(c),L}, X_{\mathcal{N}(c),R})$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu, \text{diag}(\sigma)) \longrightarrow q(z_c|X_{c,L}, X_{c,R}, X_{\mathcal{N}(c),L}, X_{\mathcal{N}(c),R}) = N(\mu_c, \mathrm{diag}(\sigma_c))$$

$$(\mu_l, \sigma_l) = \text{Decoder}_\theta(\mathbf{z}) \longrightarrow (\mu_{c,l}, \sigma_{c,l}) = \mathrm{DeepST}(z_c, z_{\mathcal{N}(c)})$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mu_l, \text{diag}(\sigma_l)) \longrightarrow p(X_c|z_c, z_{\mathcal{N}(c)}) = N(\mu_{c,l}, \mathrm{diag}(\sigma_{c,l}))$$

1. Background

2. Our Model (DeepST)

3. Results

4. Conclusion

Conclusion

1. DeepST is a deep graph convolutional network model that makes inferences on ST data.
2. DeepST addresses concerns with model selection by directly working with graphs and treating signals from neighboring cells as learnable.
3. DeepST's spatial awareness has a stronger relative prediction improvement in contrast to models that do not work on graph inputs directly.
4. For spatially independent genes, our method can select the appropriate corresponding model, avoiding spurious conclusions about spatial dependence.

Future Work

1. DeepST can naturally be extended to a graph VAE (GVAE) for better uncertainty quantification.
2. Latent features discovered by a GVAE could identify useful features that are not directly observable.
3. Reduce memory footprint of the model.