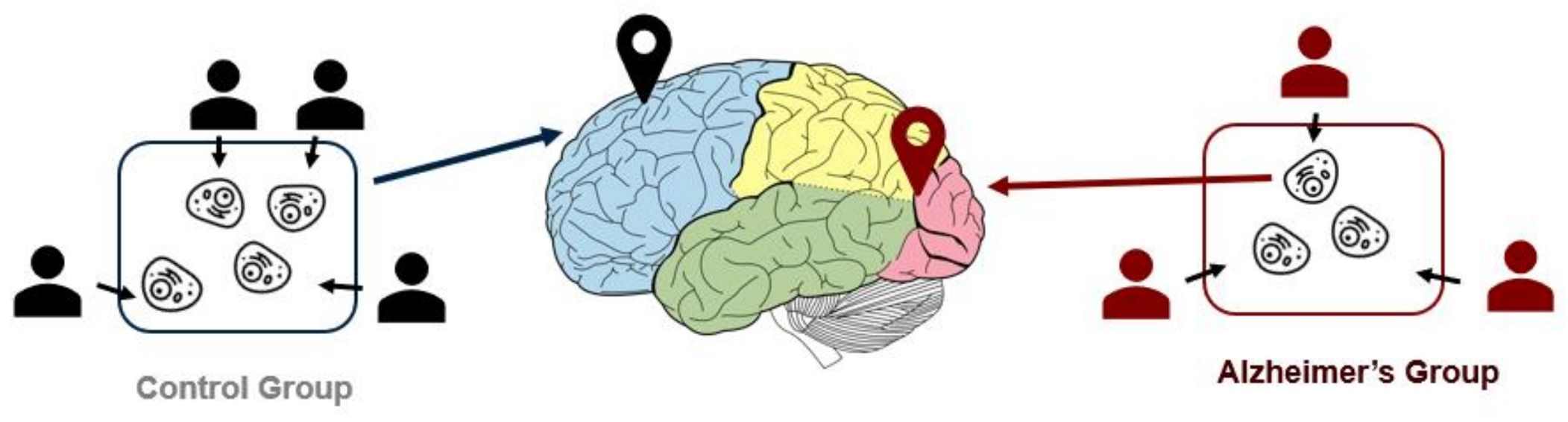
CeLEry: cell location recovery
in single-cell RNA sequencing

Qihuang Zhang Jian Hu David Dai Edward Lee Rui Xiao Mingyao Li

Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania

Introduction

Motivation: Compare the *spatial distribution* of the cells for two groups:



Research Goal:

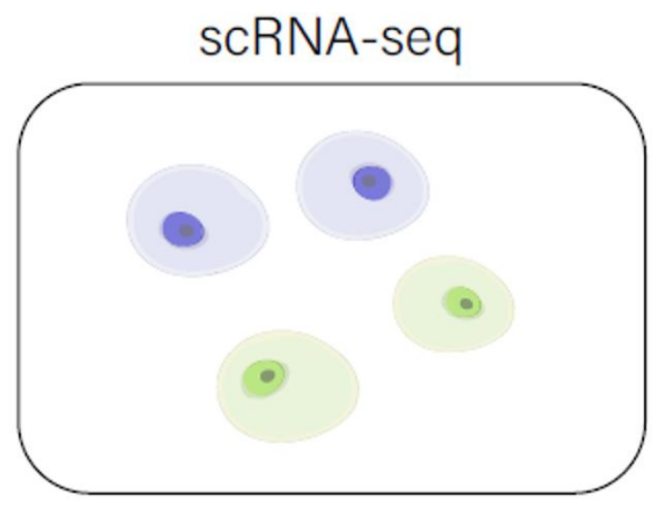
Can we predict the locations of these cells?

Data:

Query Data

✓ Gene expression

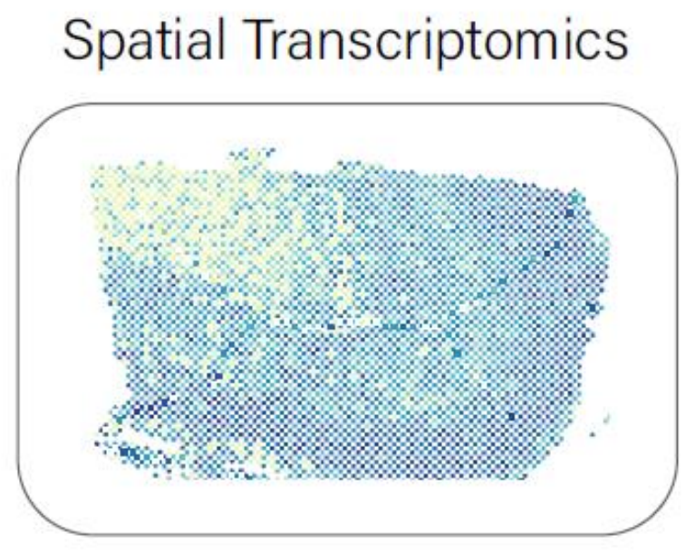
? Location



Reference Data

✓ Gene expression

✓ Location



Idea:

- scRNA data contain richer cell-level information (e.g., cell type, disease status).
- Spatial transcriptomic data have location information.
- We train a model to learn the relationship between gene expression and location and then apply it to predict the location of scRNA.

Notation

Subject:

i: a spot (reference data) or a cell (query data)

Response:

- Task 1: Coordinates Prediction
  - ***(1)*** *Point prediction*
  - ***(2)*** *Region prediction*

$$Y_i = (Y_{i1}, Y_{i2}),$$

where Y_{i1} and Y_{i2} are continuous from $[0,1]$

- Task 2: Layer Prediction

Y_i is an ordinal variable taken from $[1, 2, \dots, 7]$

Covariate

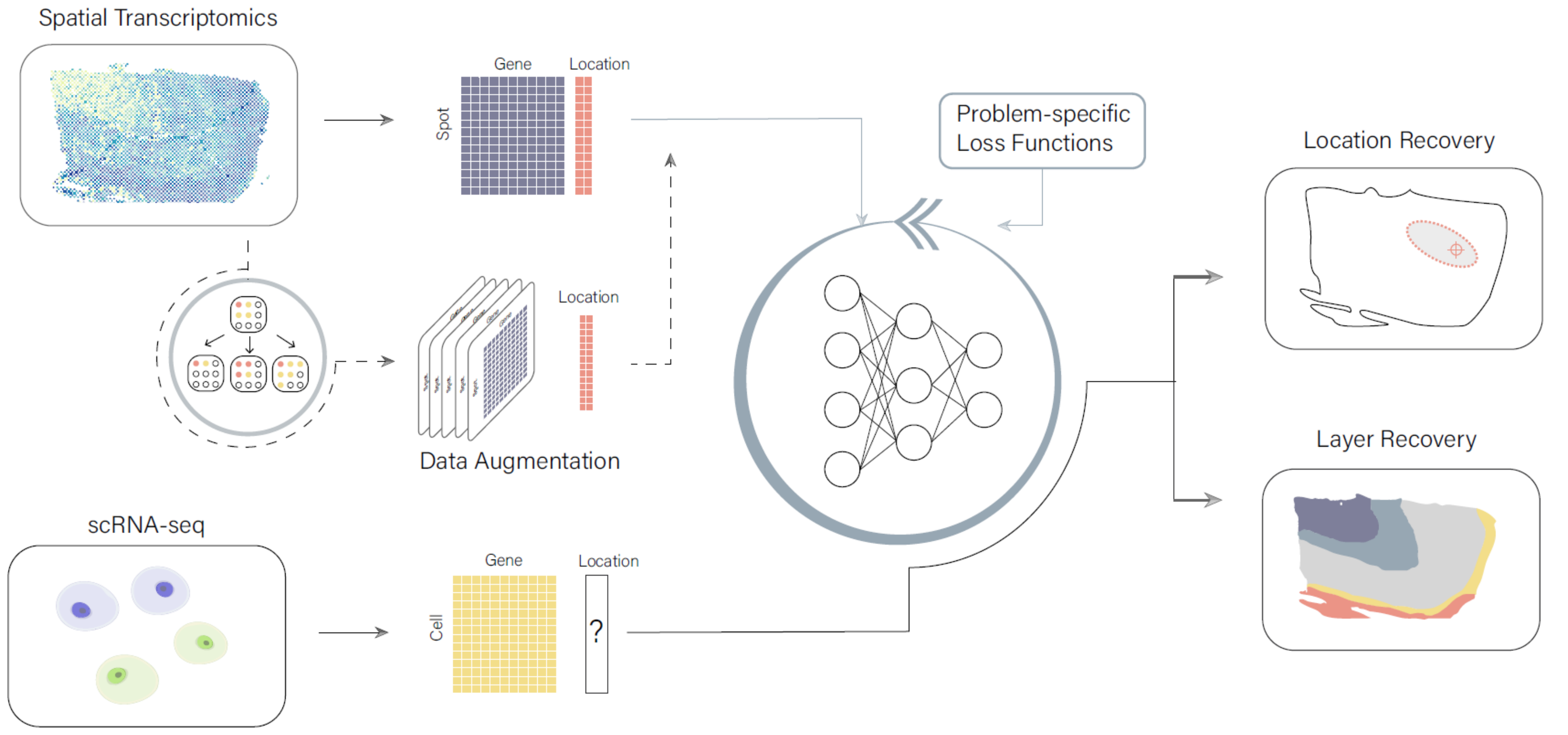
X_{ij}: Gene expression of gene j.

(A z-score normalization is performed.)

Modeling objective

Build a prediction model $f(y_i|x_i)$ to minimize the loss between the predicted value $\hat{Y}_i$ and its truth Y_i.

Method



- CeLEry takes spatial transcriptomic data as input for the training data and the scRNA-seq as testing data set.
- CeLEry optionally generates replicates of the spatial transcriptomic data via variational autoencoder then includes them as the training data together with original spatial transcriptomic data.
- A deep neural network is trained to learn the relationship between the spotwise gene expression and location information, minimizing the loss functions that are specified according to the specific problem.

Loss Functions

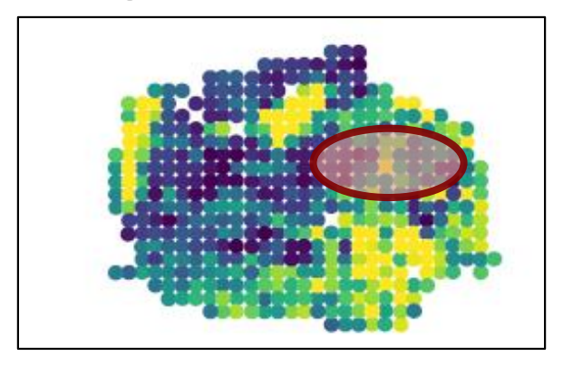
Point Prediction

(y_{i1}, y_{i2}) $(\hat{y}_{i1}, \hat{y}_{i2})$

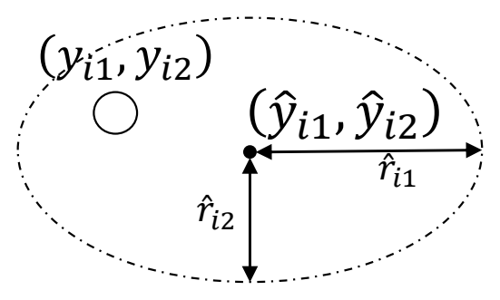
Mean squared error loss

$$\min \sum_{i}^{n} (\hat{y}_{i1} - y_{i1})^2 + (\hat{y}_{i2} - y_{i2})^2$$

Region Prediction



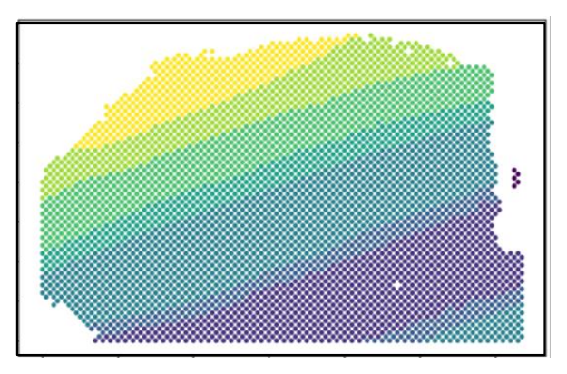
Ellipse quantile regression loss



Confidence level

$$\min \sum_{i=1}^{n} (\alpha(1-s_i) + (1-\alpha)s_i) \left| \left(\frac{y_{i1} - \hat{c}_{i1}}{\hat{r}_{i1}}\right)^2 + \left(\frac{y_{i2} - \hat{c}_{i2}}{\hat{r}_{i2}}\right)^2 - 1 \right|$$

Layer Prediction



Rank consistent ordinal regression loss

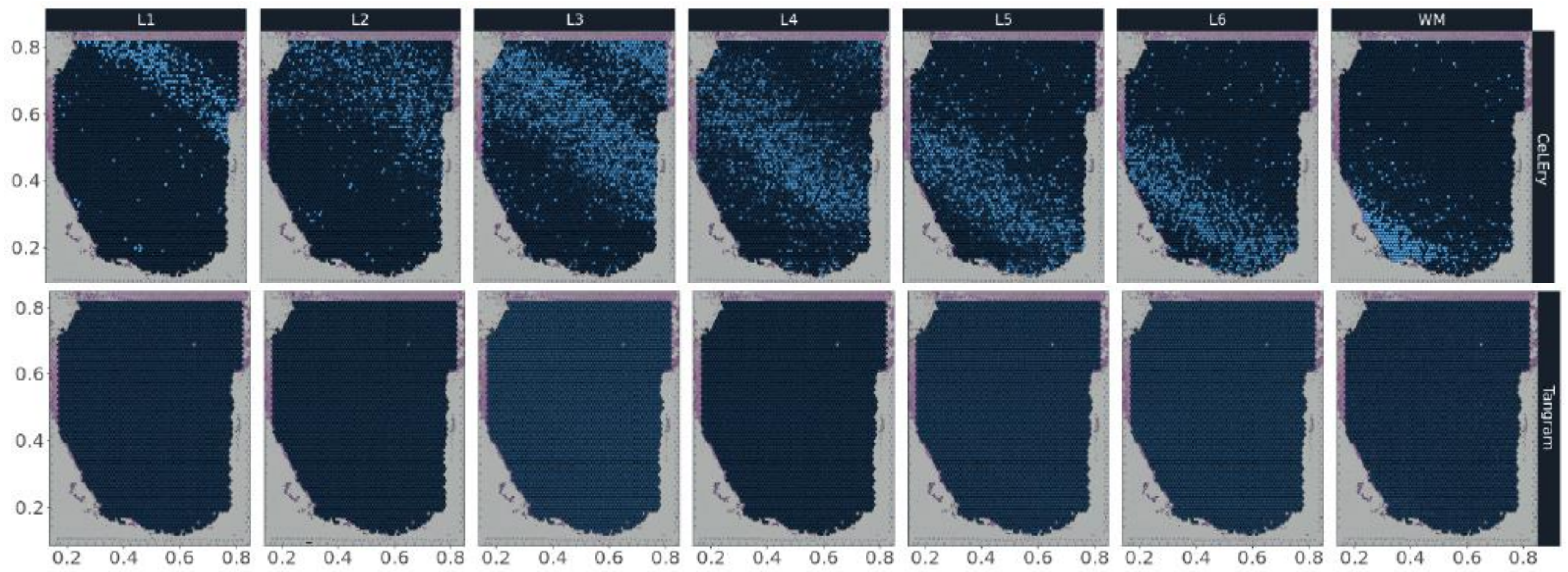
$$\sum_{i=1}^{n} \sum_{l=1}^{L-1} \log\{\sigma(\hat{a}_i + b_l)\} \cdot I(y_i > l) + \log\{1 - \sigma(\hat{a}_i + b_l)\} \cdot \{1 - I(y_i > l)\}$$

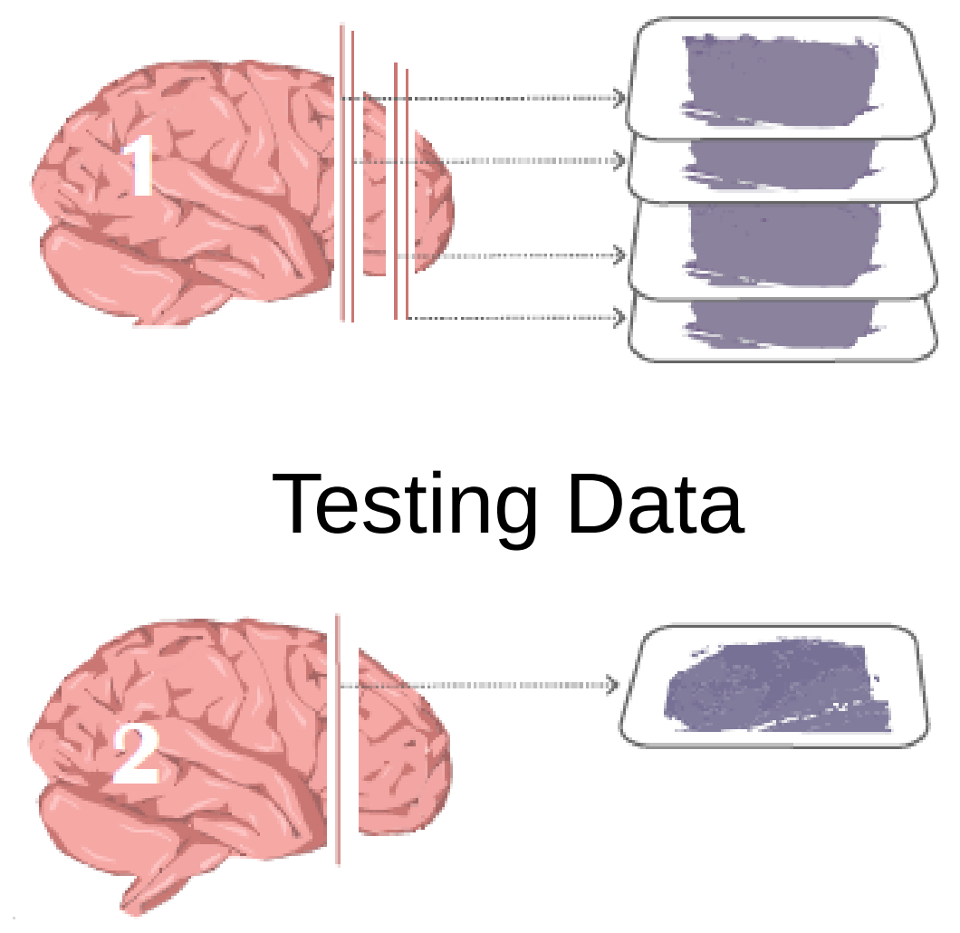
where $\sigma(x) = \frac{\exp(x_i)}{1+\exp(x_i)}$.

cut point between layer l and $l+1$

output of neural network $\hat{a}_i = f(x_i)$

Data Analyses

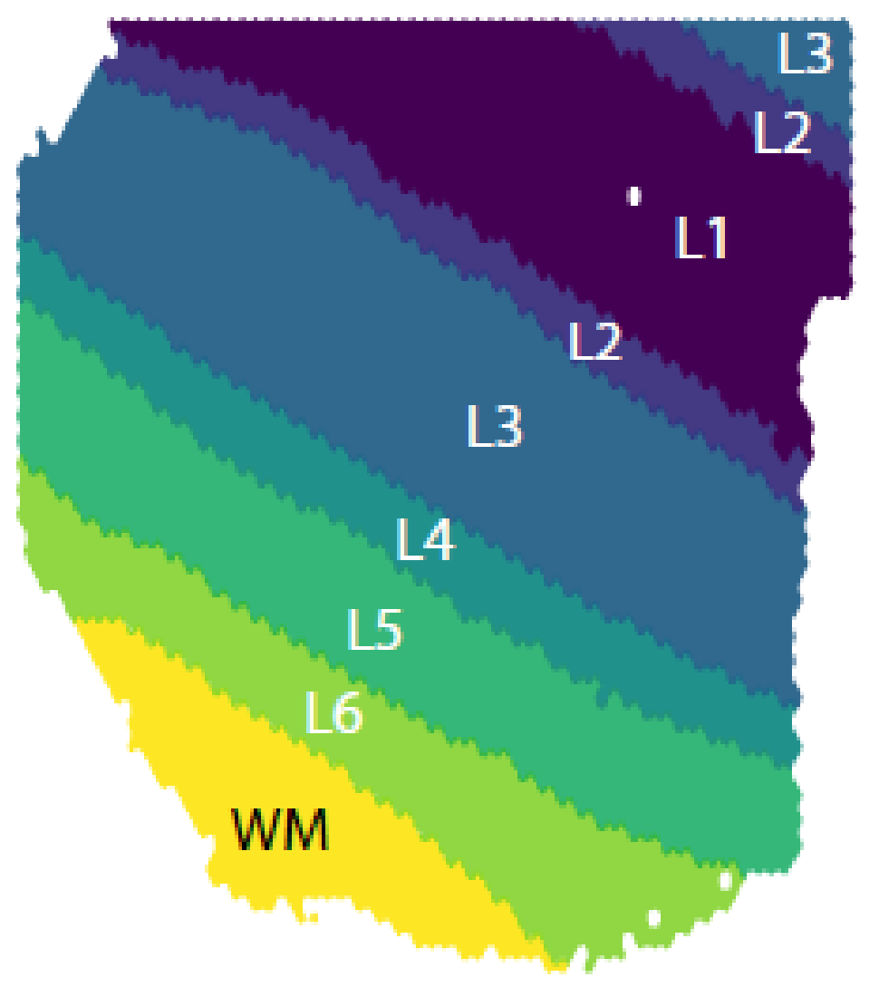




Training Data

Testing Data

Layer Truth



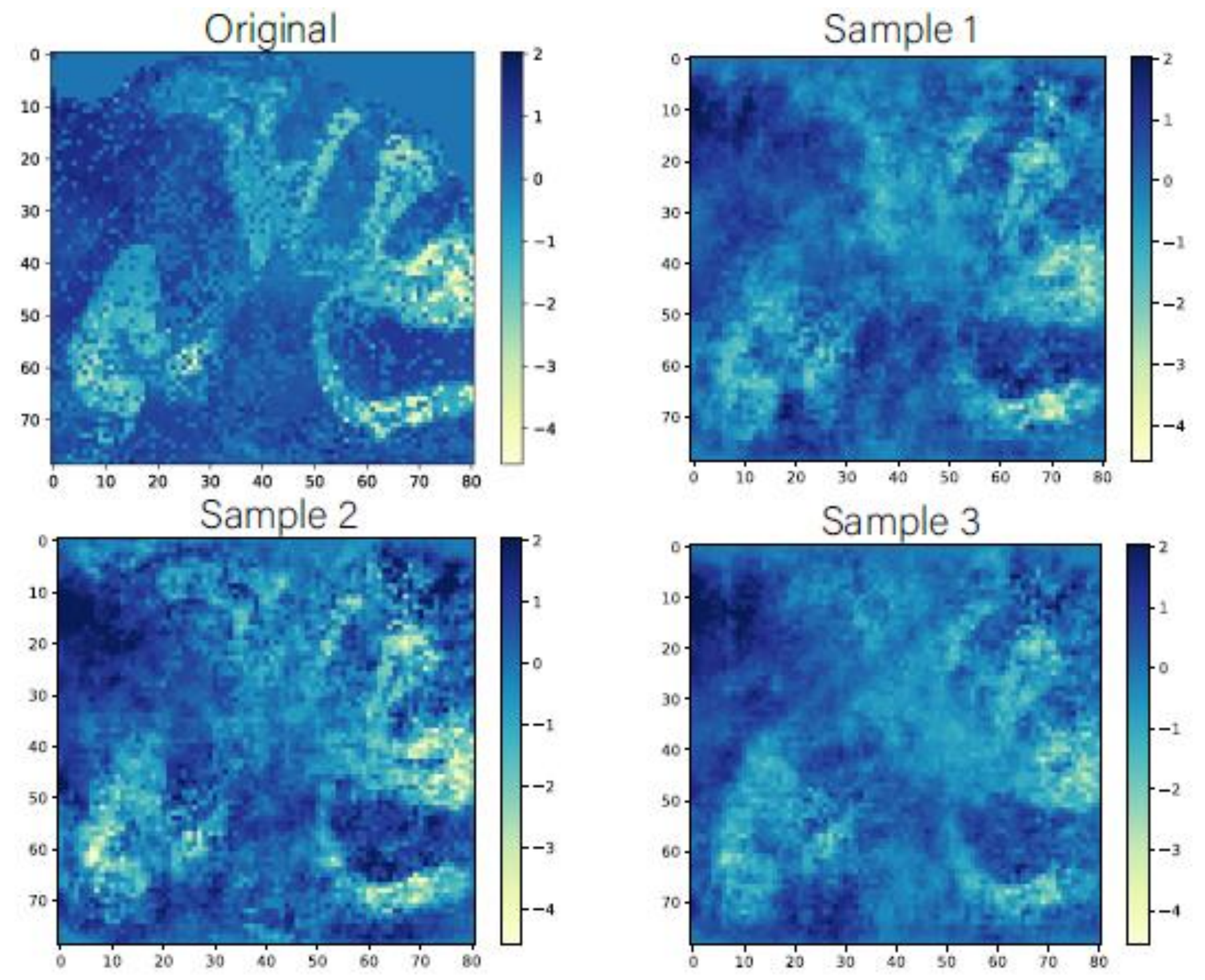
❖ Study Procedure

- We took three slices from Brain 1 to be the training data and one slice from Brain 2 to be the testing evaluation.
- For each layer, we report the probability of predicting each spot to this layer based on results from CeLEry and Tangram.
- We compare the results with the true layer segmentation.

❖ Results

- ✓ CeLEry has better accuracy in classifying the layer source of each spot.

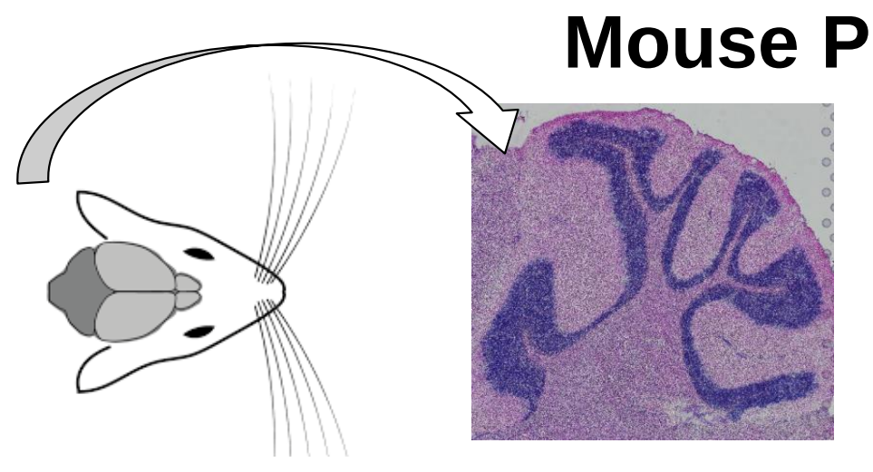
Data Augmentation



The generated samples maintained the overall pattern of the original gene map while keeping their own variation.

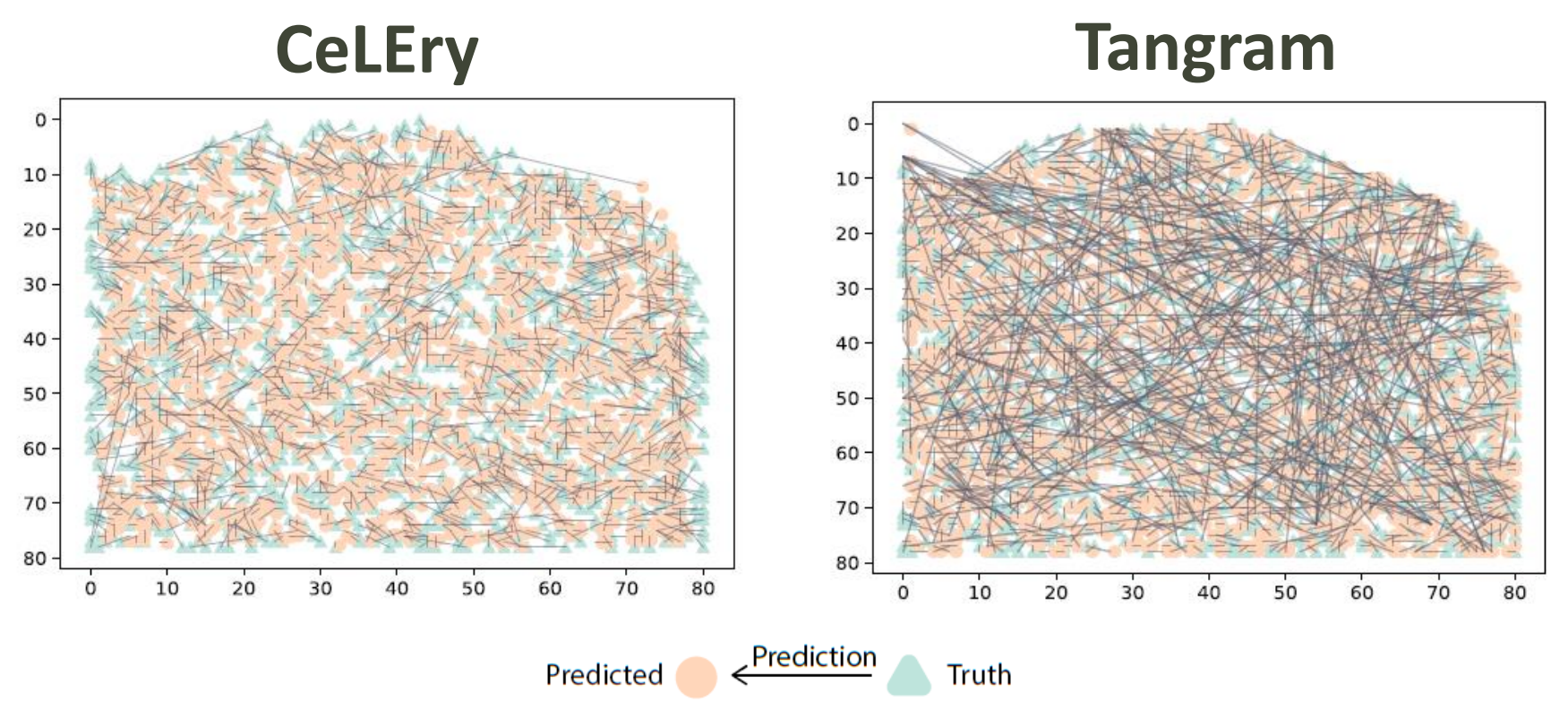
Benchmark Study

Mouse Posterior Data

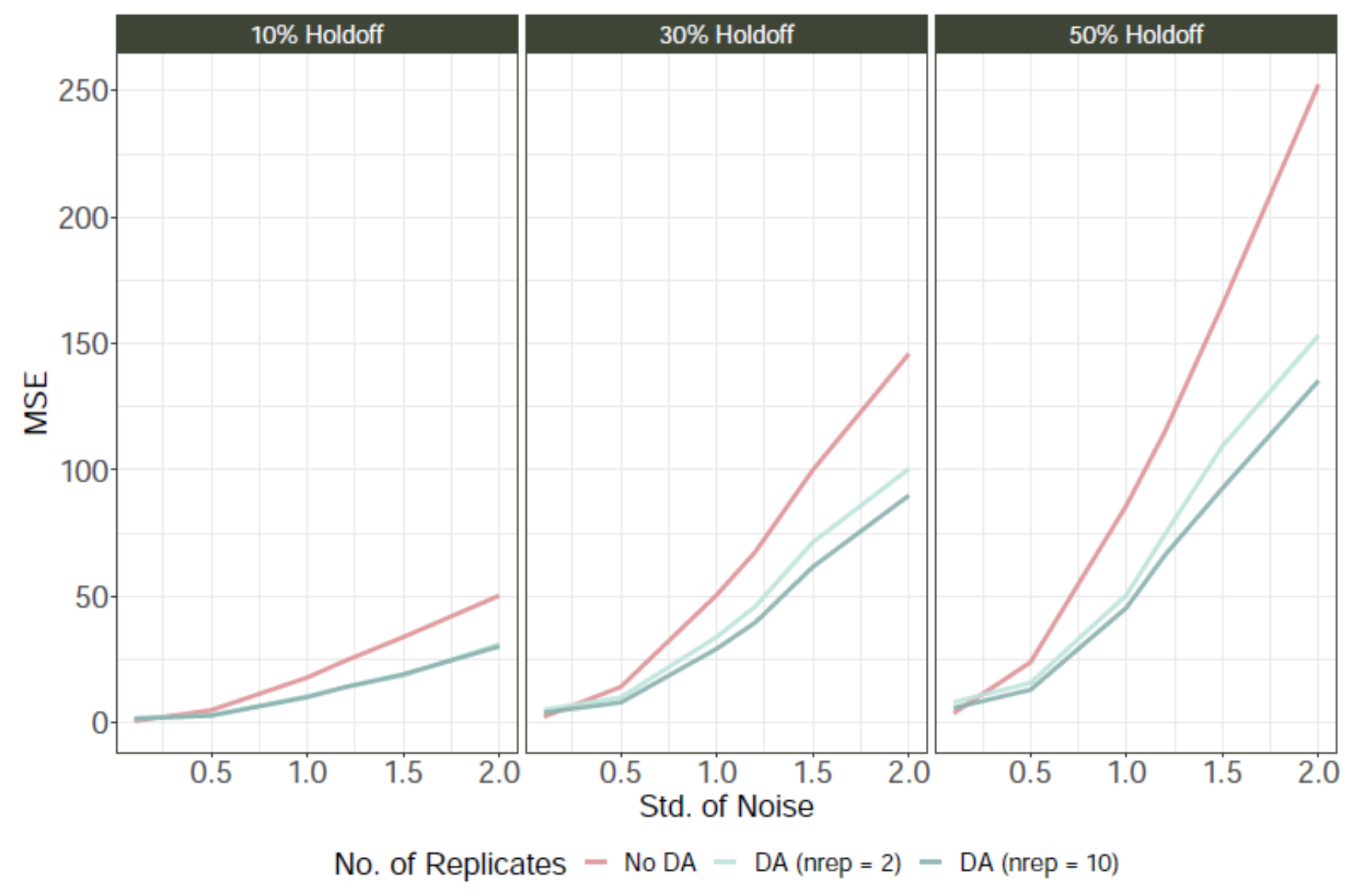


Training: 70% of spots
Testing: 30% of spots
(10%, 30%, 50%)

❖ Coordinates Prediction Accuracy

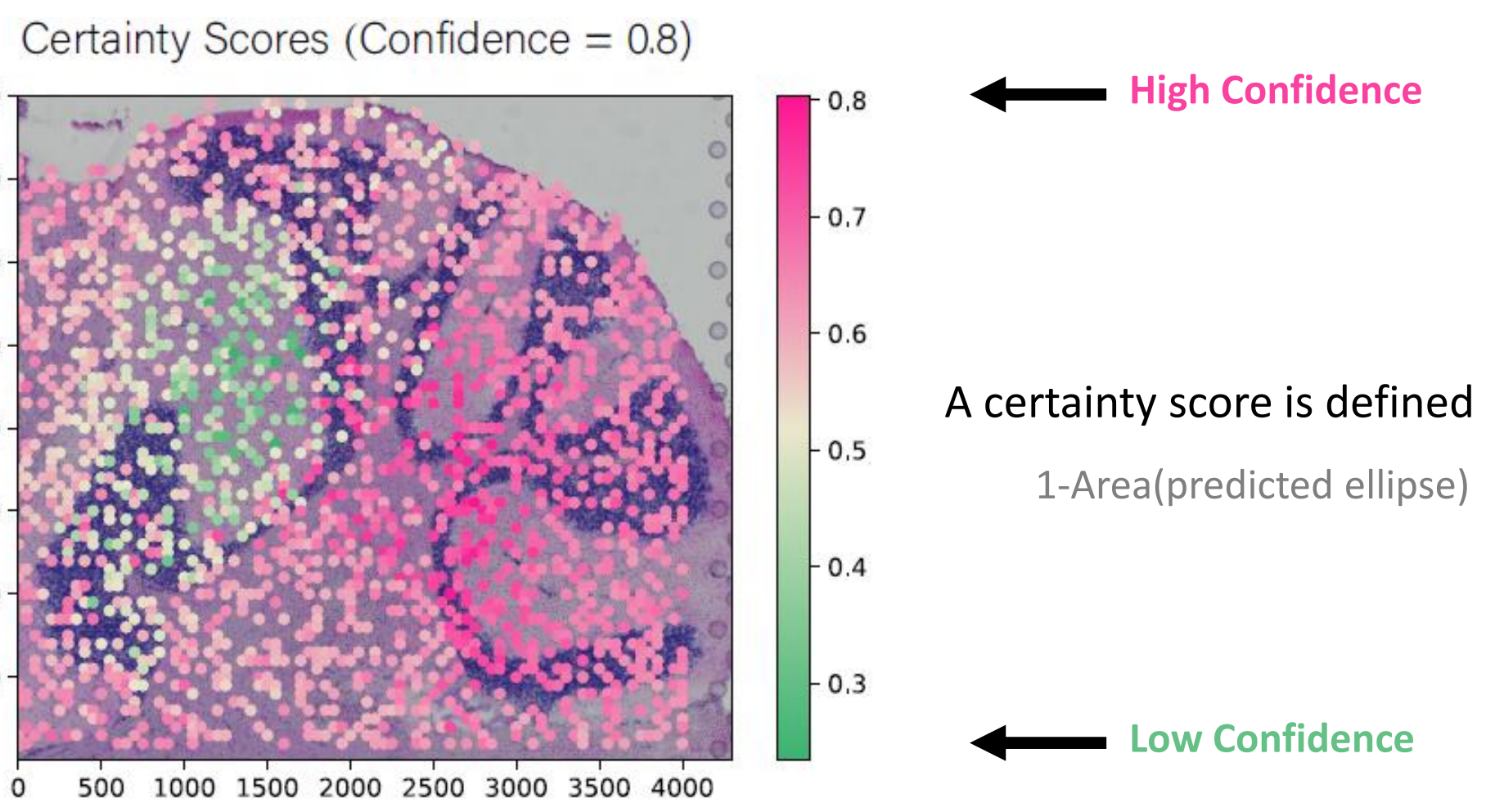


❖ Robustness against noise



The data augmentation procedure improves the robustness against the noises in the data.

❖ Prediction Uncertainty



A certainty score is defined as

1-Area(predicted ellipse)

- ✓ The high confidence spots are aligned with the dark region in the brain (granular layer)
- ✓ The low confidence spots are clustered

Acknowledgement

This project is under the supervision of Dr. Mingyao Li and Dr. Rui Xiao.