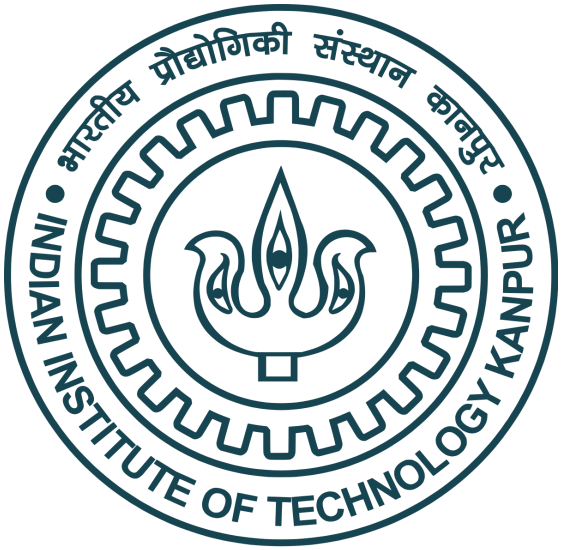
CS 690: Computational Genomics

Indian Institute of Technology, Kanpur



TEAM - 5

Authors:
Mubashshir Uddin[190516]
Suyash Mallik
Divyam Jain

Ins - Hamim Zafar

ENDSEM REPORT

Date -23/11/2023

Objective:

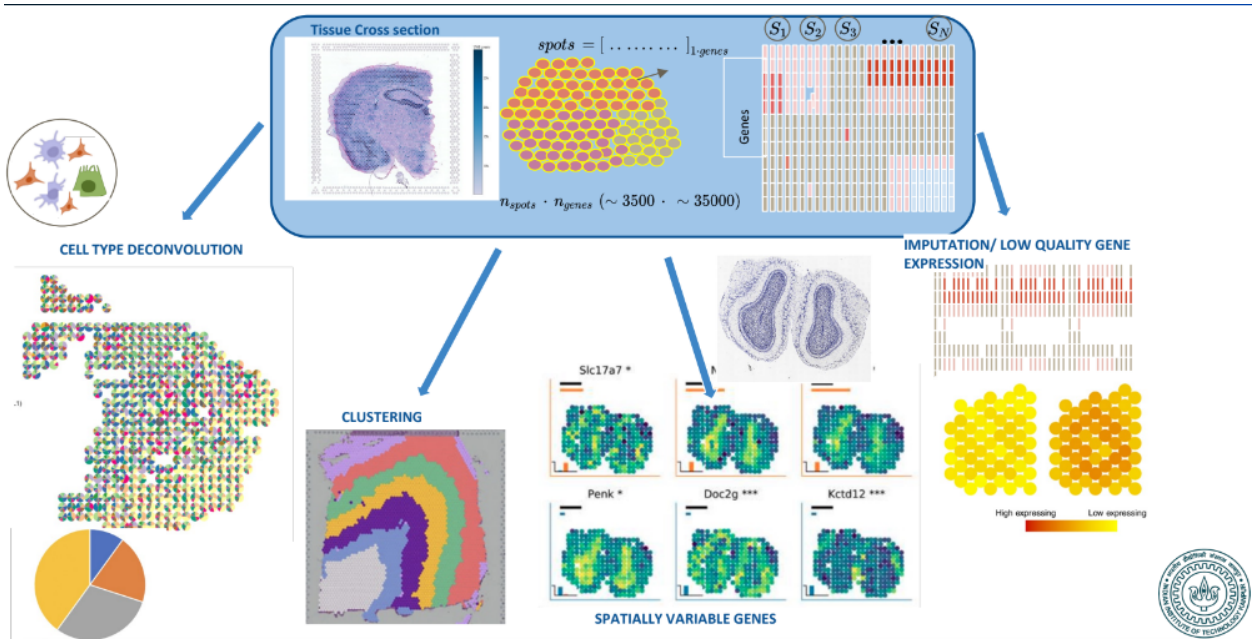
Imputation and denoising of spatial transcriptomic data

Spatially resolved transcriptomics (SRT) provides gene expression close to, or even superior to, single-cell resolution while retaining the physical locations of sequencing and often also providing matched pathology images. However, SRT expression data suffer from high noise levels, due to the shallow coverage in each sequencing unit and the extra experimental steps required to preserve the locations of sequencing.

The goal of this project is to develop a deep generative model for spatial transcriptomics data that will utilize the information from the physical locations of sequencing, and the tissue organization reflected in corresponding pathology images to learn a latent representation of the data which can be used for denoising the data. Several graph neural network models are available for data imputation. Some of these models can be extended to incorporate the spatial graph for imputation. Standard datasets used by other imputation methods can be used for evaluation.

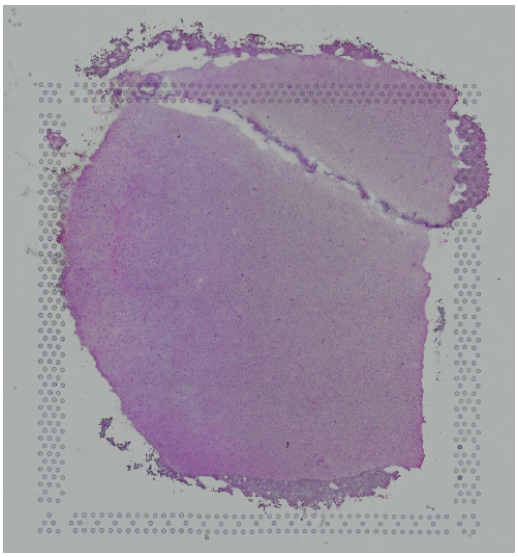
Introduction:

Spatially transcriptomics refers to the mutual compilation of data in the form of microscopic images and in-place gene sequencing:

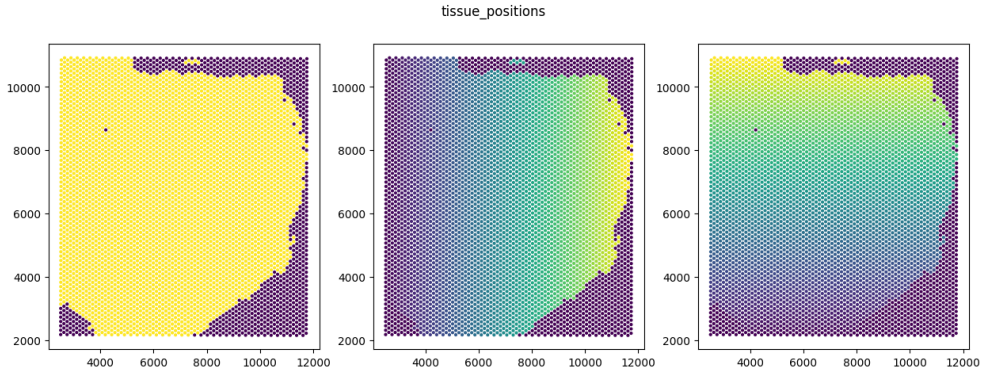


Modalities:

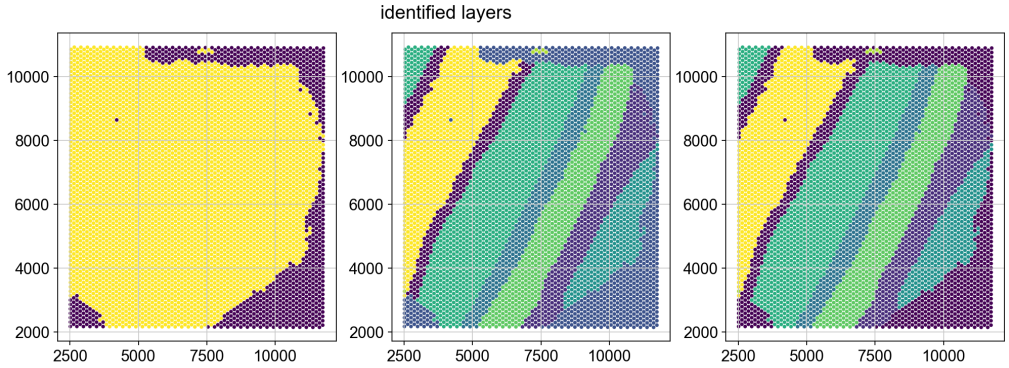
- Image Modality:
  - ○ High-resolution microscopic image of the tissue with a scale factor of 0.15
  - ○ Passing in the image through convolutional VAE, turning into a latent vector
  - ○ Scale-invariant VAE training to best reproduce the image implies that most of the data from the image is retained in the latent space.
  - ○ Passing image in the forms of different scale patches, e.g. if the image is 100 x 100 then we pass 100,00 (1x1), 2500(2x2), 625(4x4), .., 4(50x50),1(100x100). Set of image patches in a hyperparametric combination, letting us control the information sharpness content of the image at different levels.



- Spatial position modality:
  - ○ The spatial positions of the spots that have been mapped are provided -



  - ○ Above is the information on the gradient space in the cells and the presence /absence of the spots.
  - ○ We also have available metadata.tsv file which is downloaded from the 10xvisium database, has information about which layer the given barcode comes from, we have heavily mixed but nearly complete barcode mapping from the metadata.tsv to the tissue_positions_list.csv.

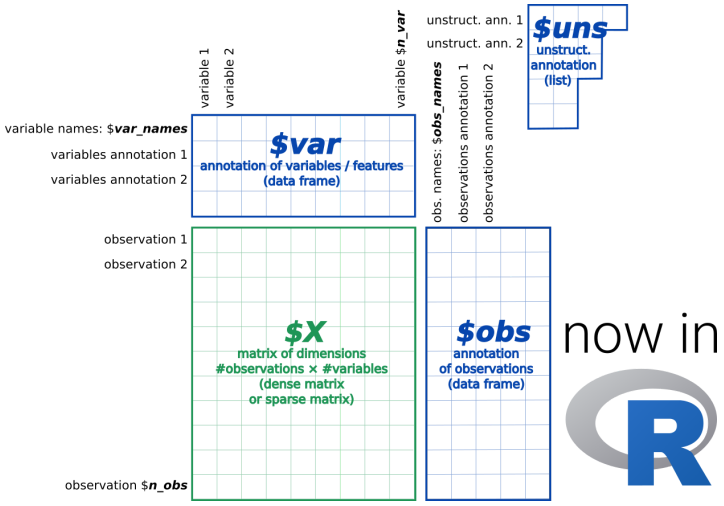


The data presenting tissue positions list - x,y are the coordinates of the barcode

	barcode	present	xgrad	ygrad	x	y	layer_guess	color
0	ACGCCTGACACGCGCT-1	0	0	0	2510	2174	NaN	2
1	TACCGATCCAACACTT-1	0	1	1	2630	2243	NaN	2
2	ATTAAAGCGGACGAGC-1	0	0	2	2511	2312	NaN	2
3	GATAAGGGACGATTAG-1	0	1	3	2631	2381	NaN	2
4	GTGCAAATCACCAATA-1	0	0	4	2511	2450	NaN	2

(4992, 8)

- Single-cell RNA sequencing data:
  - A relatively lower quality sc-RNA seq read is obtained from each one of the labelled cells -




```
ADATA uniqlyr: {'BA9', 'BA24', 'BA46', 'PFC'}

Adata obs:
                                    ground_truth batch             barcode
GAGGTGAAGTGCGTGA-1_4899_BA24-0   Neu-NRGN-II     0  GAGGTGAAGTGCGTGA-1
AGAGCGAAGCACCGTC-1_5554_BA24-0   Neu-NRGN-II     0  AGAGCGAAGCACCGTC-1
CACCAGGGTGTATGGG-1_5841_BA9-0    Neu-NRGN-II     0  CACCAGGGTGTATGGG-1
CAGAGAGTCAGTTTGG-1_5841_BA9-0    Neu-NRGN-II     0  CAGAGAGTCAGTTTGG-1
AGTGAGGTCATACGGT-1_5387_BA9-0    Neu-NRGN-II     0  AGTGAGGTCATACGGT-1
...                                      ...   ...                ...
GCCAAATCACCAGTTA-1_6032_BA24-16       AST-PP    16  GCCAAATCACCAGTTA-1
AAAGATGGTACCGTTA-1_5945_PFC-16        AST-PP    16  AAAGATGGTACCGTTA-1
CCGTACTTCACGATGT-1_5531_BA24-16       AST-PP    16  CCGTACTTCACGATGT-1
CACAGGCTCTGGAGCC-1_5565_BA9-16        AST-PP    16  CACAGGCTCTGGAGCC-1
GACGGCTGTTCCTCCA-1_5939_BA9-16        AST-PP    16  GACGGCTGTTCCTCCA-1
```

Adata Obs: [*20904 rows x 5 columns*]

Ground truth layers (Obs):
['Oligodendrocytes', 'Endothelial', 'IN-PV', 'Neu-mat', 'IN-VIP', 'OPC', 'AST-FB', 'L5/6-CC', 'L2/3', 'L4', 'AST-PP', 'Neu-NRGN-I', 'L5/6', 'Neu-NRGN-II', 'Microglia', 'IN-SV2C', 'IN-SST']

Adata vars(gene names){*59074 rows x 0 columns*}: [DDX11L1, WASH7P, MIR6859-3, RP11-34P13.3, MIR1302-9, FAM138A, OR4G4P, OR4G11P, Etc.,...]

```
AnnData object with n_obs × n_vars = 20904 × 59074
    obs: 'ground_truth', 'batch'
{'L2', 'L6', nan, 'L4', 'WM', 'L3', 'L5', '-1', 'L1'}
{0, 1, 2, 3, 4, 5, 6, 7, 8}
{0, 1, 2, 3, 4, 5, 6, 7, -1}
4992
```

Denoising algorithms proposed:

Learn the best predictions using the latent. then check the variance from the best predictor, how if the variance is greater than a threshold then impute.