Week 12:
Learning from multi-modal / multi-view data

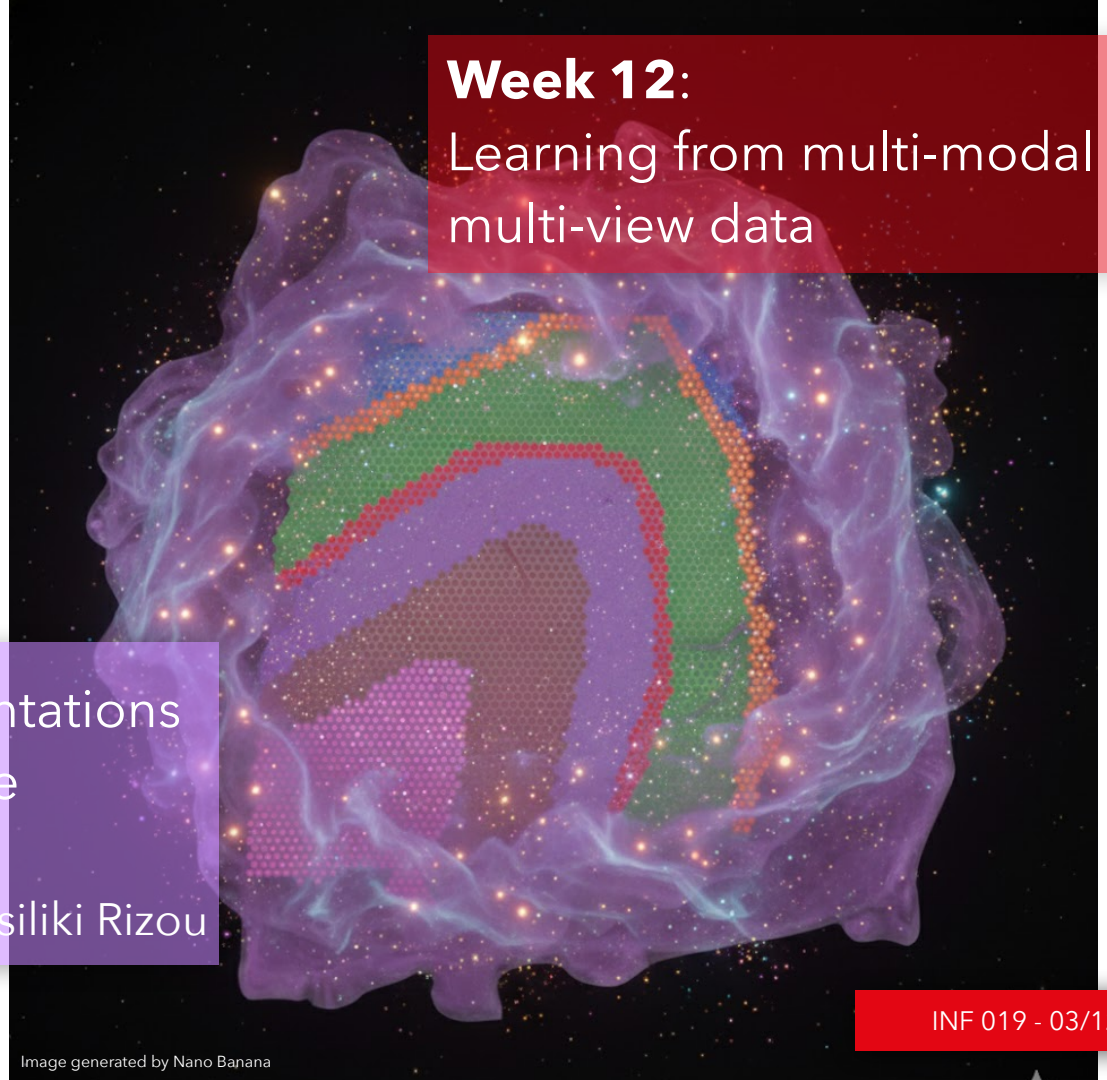


Image generated by Nano Banana

EE-626 - Graph representations for biology and medicine

presented by Vasiliki Rizou

INF 019 - 03/12/25

Who am I?

"Disentanglement in Multimodal Learning"

The use of multiple modes of communication (e.g., text, images sound, video) to enhance understanding.

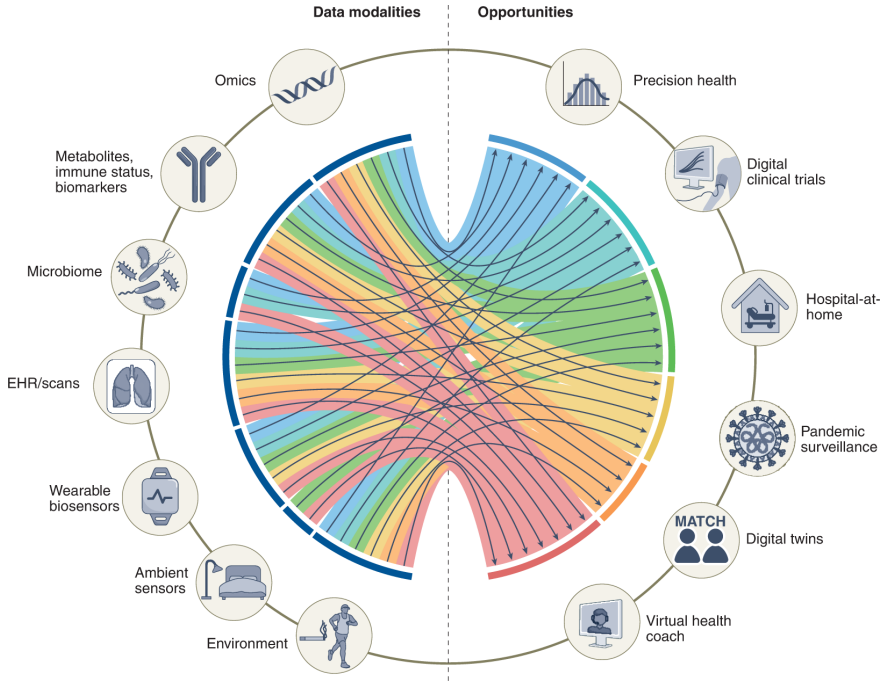


Image from [1]

1^{st} year PhD, EDEE

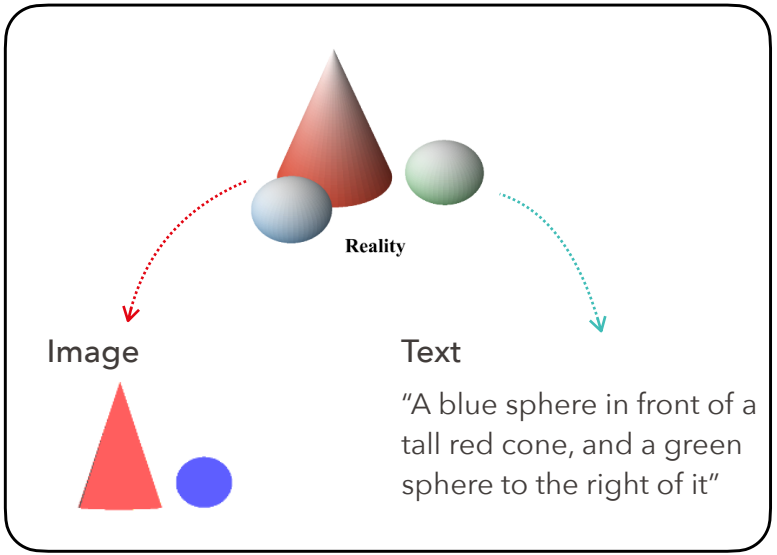
Lab: LTS4

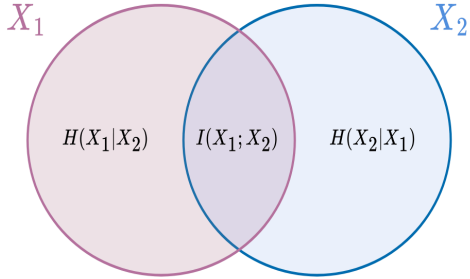
Supervisors:
Dr. Dorina Thanou,
Prof. Pascal Frossard

[1] Julian N. Acosta et. al., "Multimodal biomedical AI", Nature 2022

Who am I?

"Disentanglement in Multimodal Learning"





Goal: Extract the unique z_1, z_2 and the mutual information component z_{12}

Why?

- Interpretability
- Handling of missing modalities
- Generating missing modalities
- Going beyond the shared information component

nature communications

Article

https://doi.org/10.1038/s41467-025-56276-0

STAIG: Spatial transcriptomics analysis via image-aided graph contrastive learning for domain exploration and alignment-free integration

Received: 3 January 2024

Accepted: 6 January 2025

Yitao Yang[1], Yang Cui[1], Xin Zeng[1], Yubo Zhang[1], Martin Loza[2], Sung-Joon Park[2] & Kenta Nakai[1,2]

Spatial Transcriptomics (ST)

Spatial Transcriptomics is a method to map and quantify gene expression (mRNA transcripts), preserving the native location within the cells or tissues. Links molecular function directly to anatomical structure.

Imaging - based

Targeted/biased
Pre-selected targets

Subsecullar resolution
[higher resolution]

e.g., MERFISH

Sequencing - based

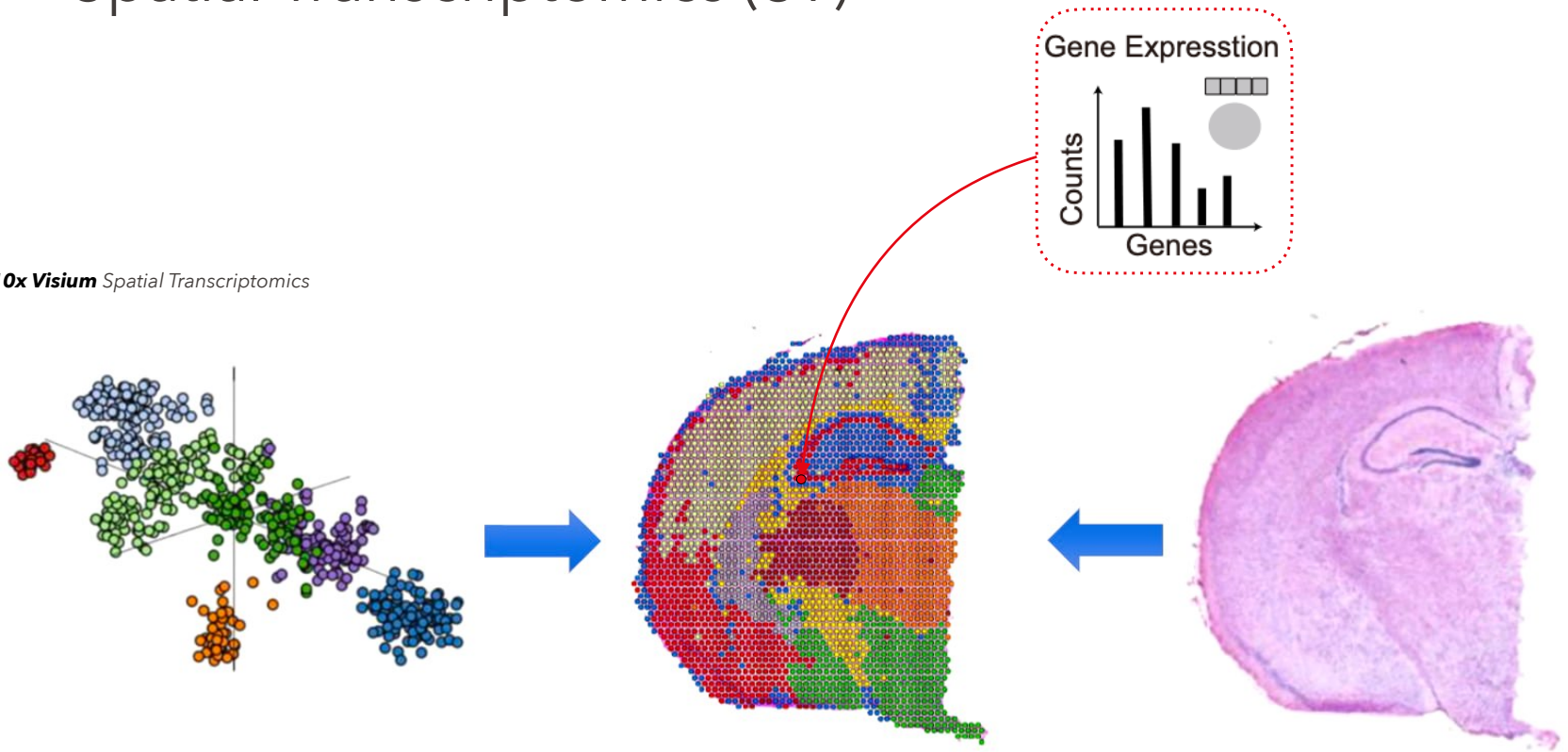
Untargeted / unbiased
Whole transcriptome
[higher throughput]

Lacks singles cell res

e.g., 10x Visium

Spatial Transcriptomics (ST)

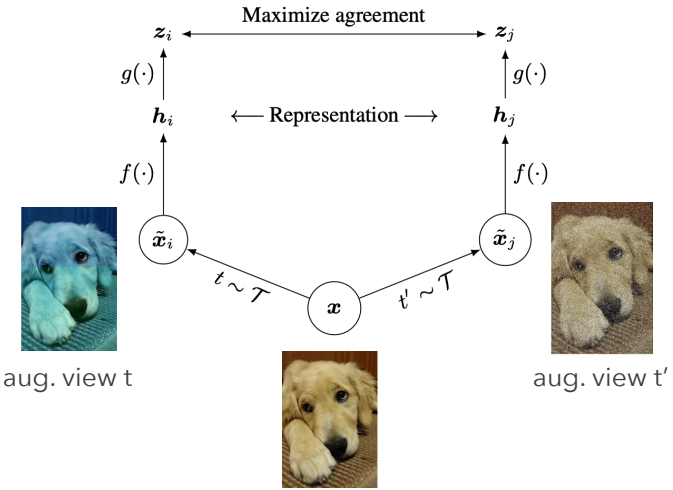
10x Visium *Spatial Transcriptomics*

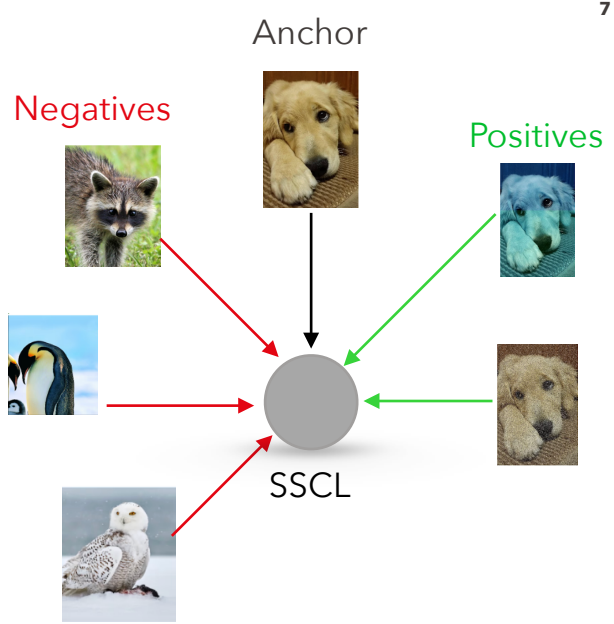


Adapted from 10x Genomics

Contrastive Learning [1]

By leveraging similarity and dissimilarity, contrastive learning enables models to map **similar instances close together** in a latent space while **pushing apart those that are dissimilar.**





Given: $\mathcal{X} = \{x, x^+, x_1^-, x_2^-, \ldots, x_N^-\}$ a similarity function $s(.)$

Goal: $s(f(x), f(x^+)) \gg s(f(x), f(x_i^-)), \forall i, \in \{1, \ldots, N\}$

InfoNCE loss:

$$\mathcal{L}_{\mathcal{N}} = -\mathbb{E}_{\mathcal{X}}[\log \frac{\exp(s(f(x), f(x^+)))}{(s(f(x), f(x^+))) + \sum_{j=1}^{j} (s(f(x), f(x_j^-)))}]$$

- [1] Google Brain, Ting Chen et. al. ,"A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Challenges & Motivation

Goal: Identification of spatial regions using spatial transcriptomics

STAIG: Using image-guided graph contrastive learning to analyze spatial transcriptomics.

Challenges in current ST methods:

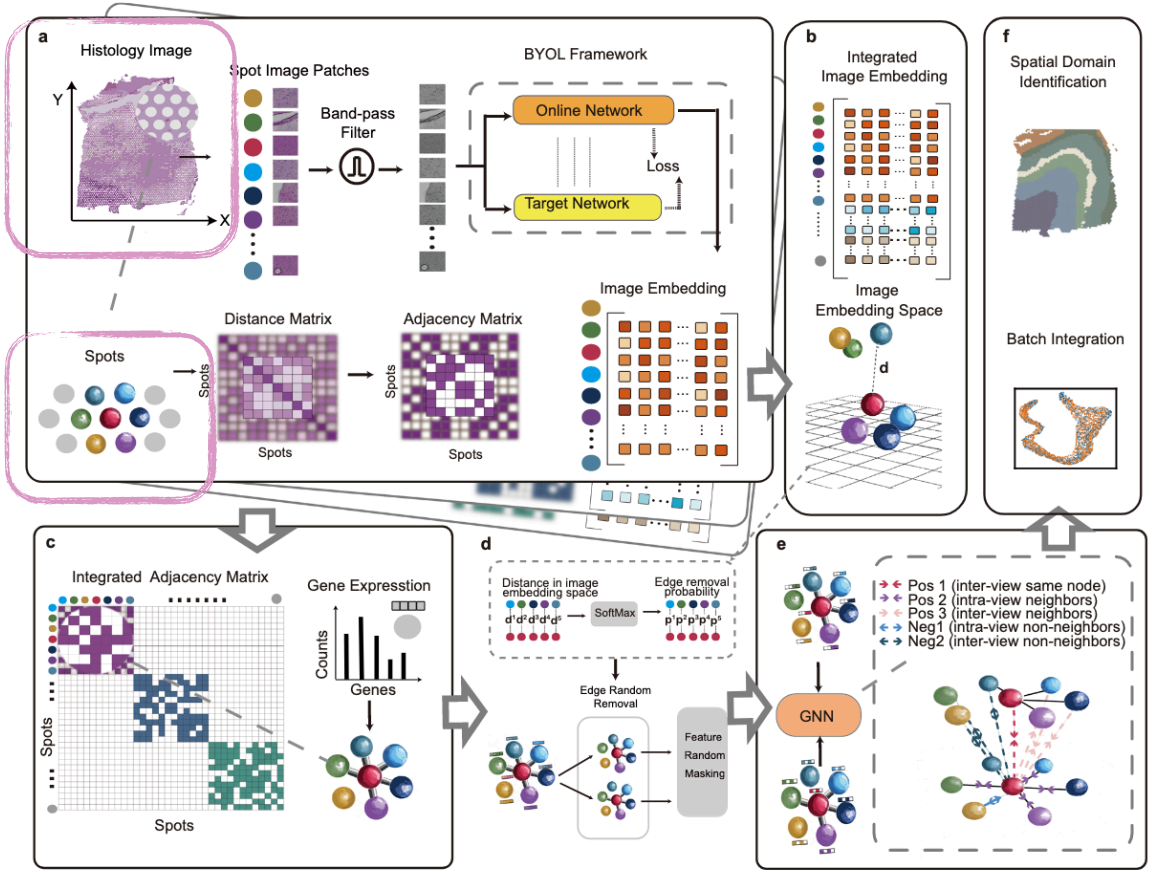
- Biases in static graph structure
- Limitations in integrating histological images
- Need of extra alignment tools for multiple slices

Contributions of **STAIG**:

- Extracts valid information from images without pre-training.
- Integrated gene expression, spatial data, and histological images.
- Dynamically adjusts graph structures.
- Enables end-to-end batch integration without pre-alignment.

Method overview

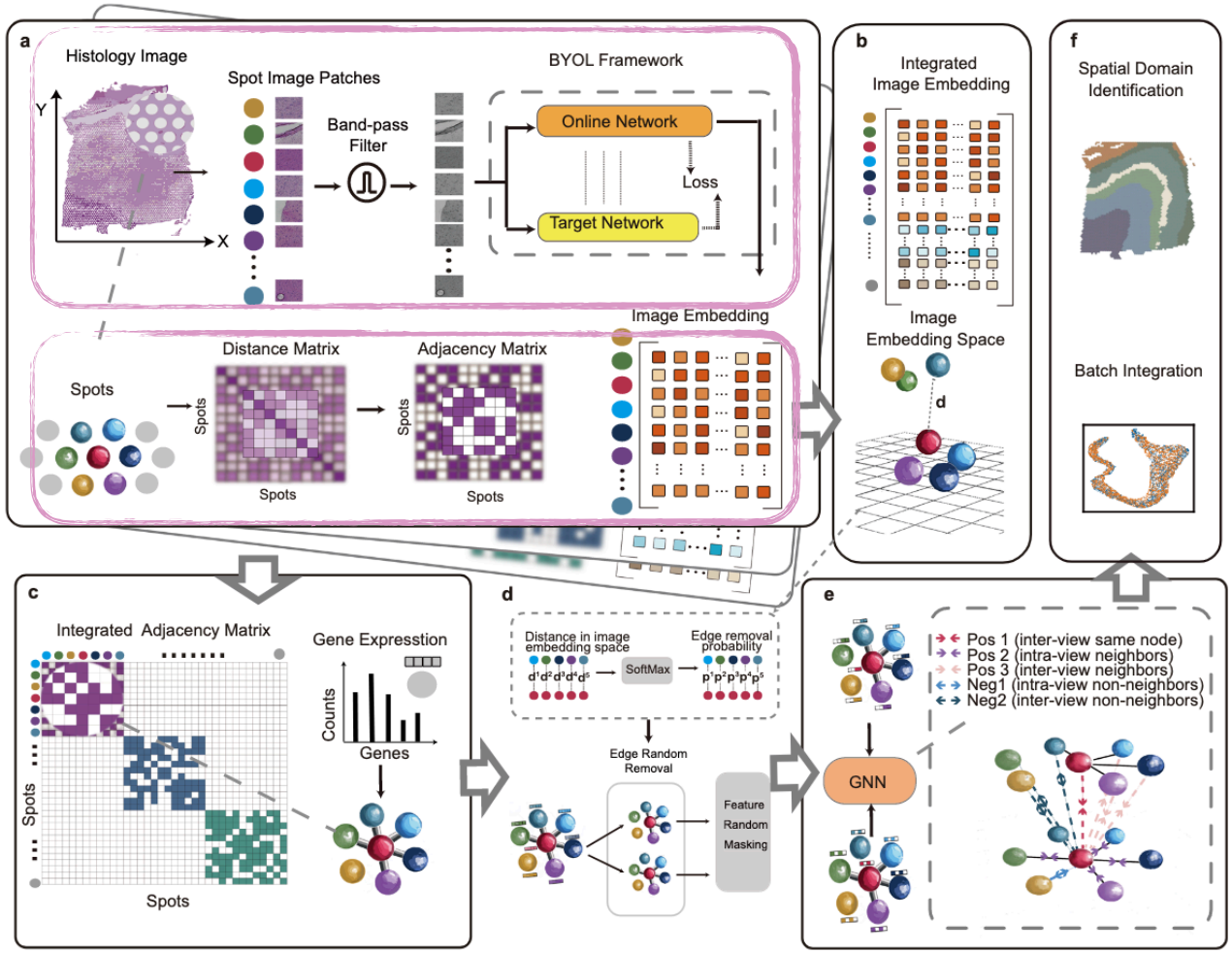
ST data include **tissue slices with spots** (spatial coords + gene counts) and, when available, **H&E images**.



Method overview

- Images are segmented into patches that align with the locations of the spots. Then each patch undergoes denoising.
- Then each patch is processed by a BYOL network $\Longrightarrow$ patch level embeddings. $C \in \mathbb{R}^{Nx64}$, each row $c_i \rightarrow$ image features for spot v_i

- For each slice basic graph construction $G = (V, E)$, V: set of N spots, E: edges derived with KNN.
- Extracting $X \in \mathbb{R}^{NxF} \rightarrow$ gene expression matrix



Method overview

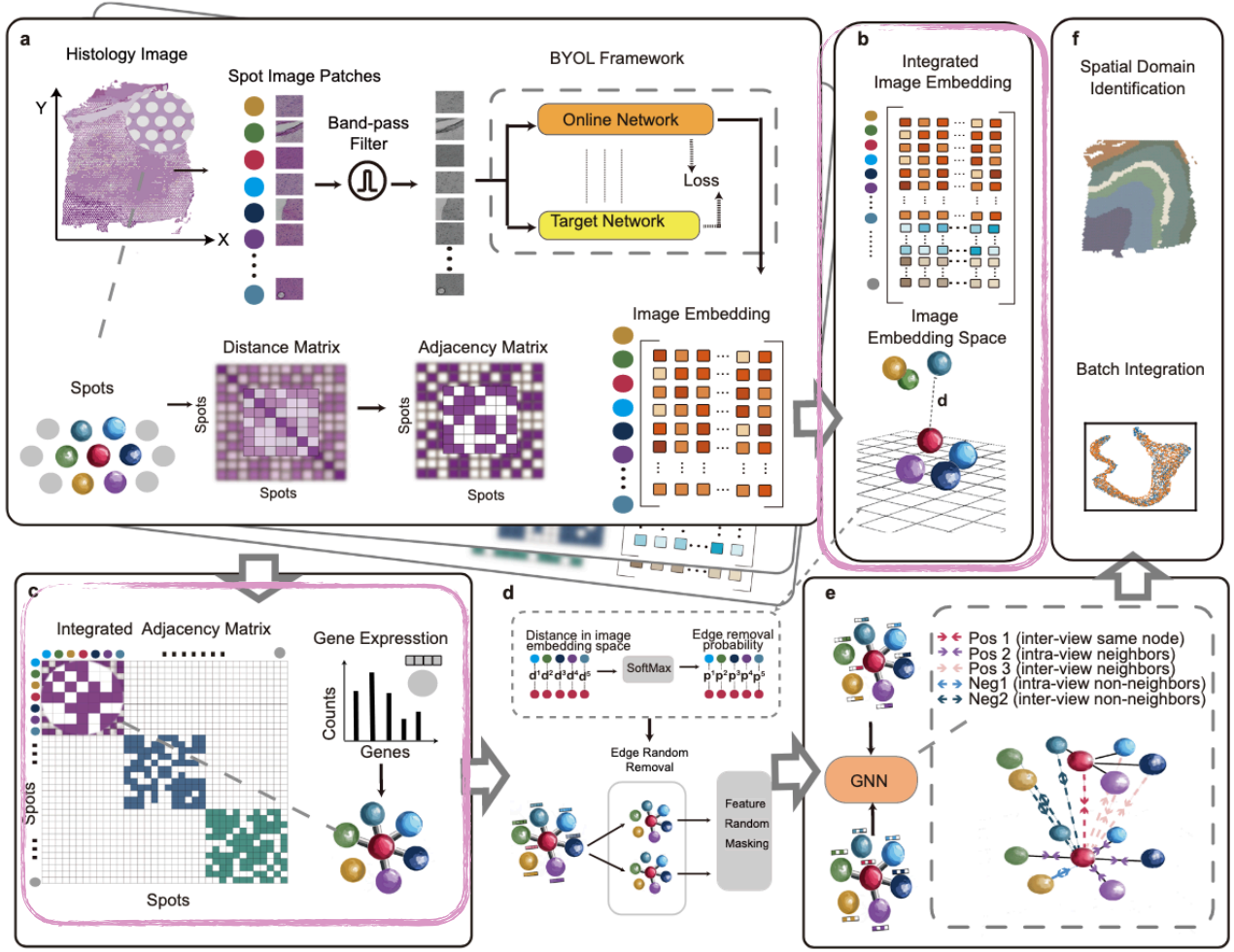
- Given m slices with respective number of slots $N_1, N_2, \ldots, N_m$ the total spot count is T.

- The respective feature images $C^1, C^2, \ldots, C^m$ are concatenated vertically

$$C = \begin{bmatrix} C^1 \\ \cdots \\ C^m \end{bmatrix} \in \mathbb{R}^{Tx64}$$

- The adjacency matrices $A^1, \ldots, A^m$ are combined diagonally

$$A = \begin{bmatrix} A^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A^m \end{bmatrix} \in \mathbb{R}^{TxT}$$

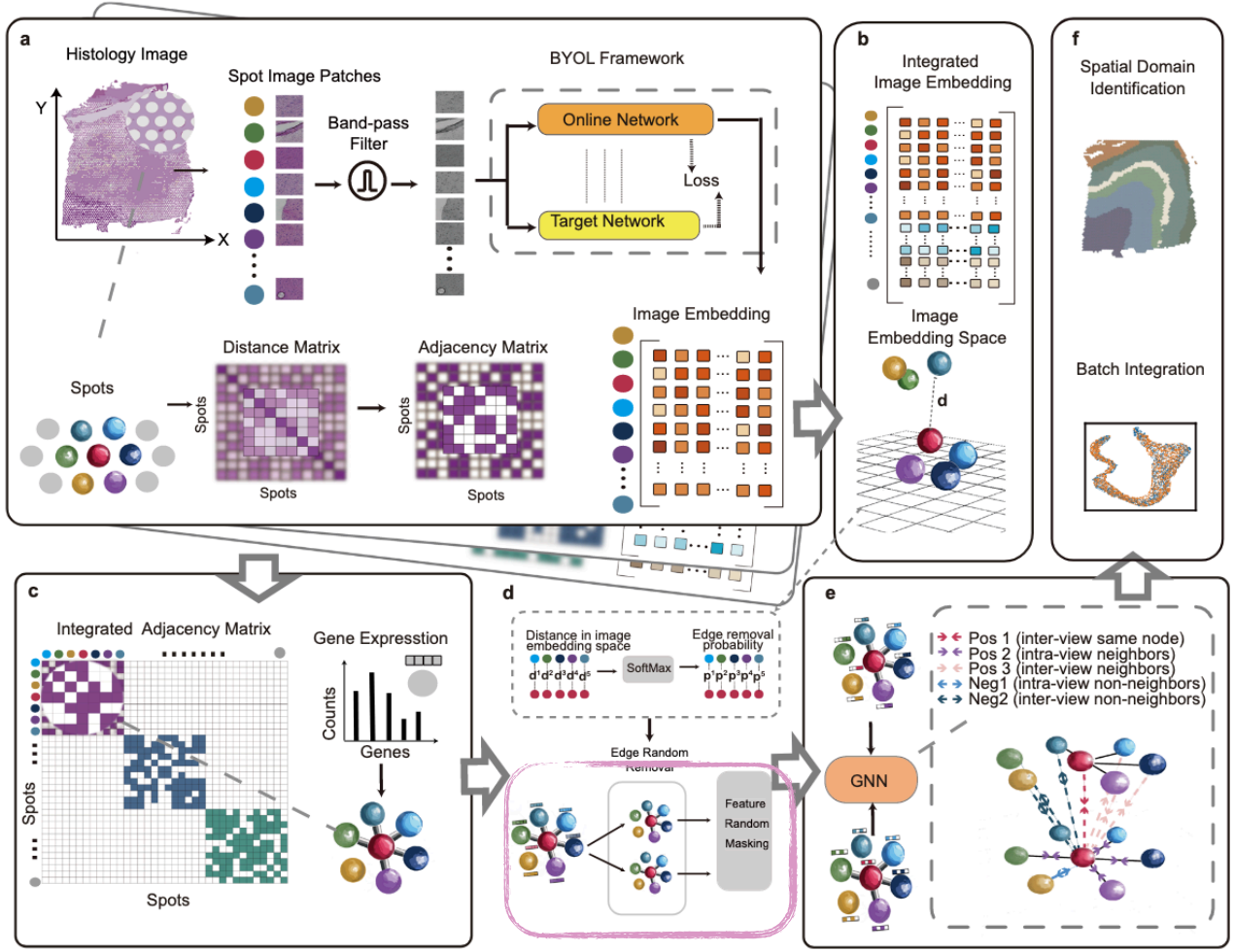


Method overview

- Generate two augmented views G_1, G_2 from A by edge removal from G:

1. Given C the image spatial matrix $D^{img} \in \mathbb{R}^{TxT}$ is computed ($D_{ij}^{img} \rightarrow$ distance between v_i, v_j)
2. Obtain probability matrix $P = softmax(D^{img}) \in \mathbb{R}^{TxT}$

- The associated modified gene expressions $\mathcal{X}_1, \mathcal{X}_2 \rightarrow$ random masking of gene features



Method overview

- The GNN processes the two views G^1, G^2:

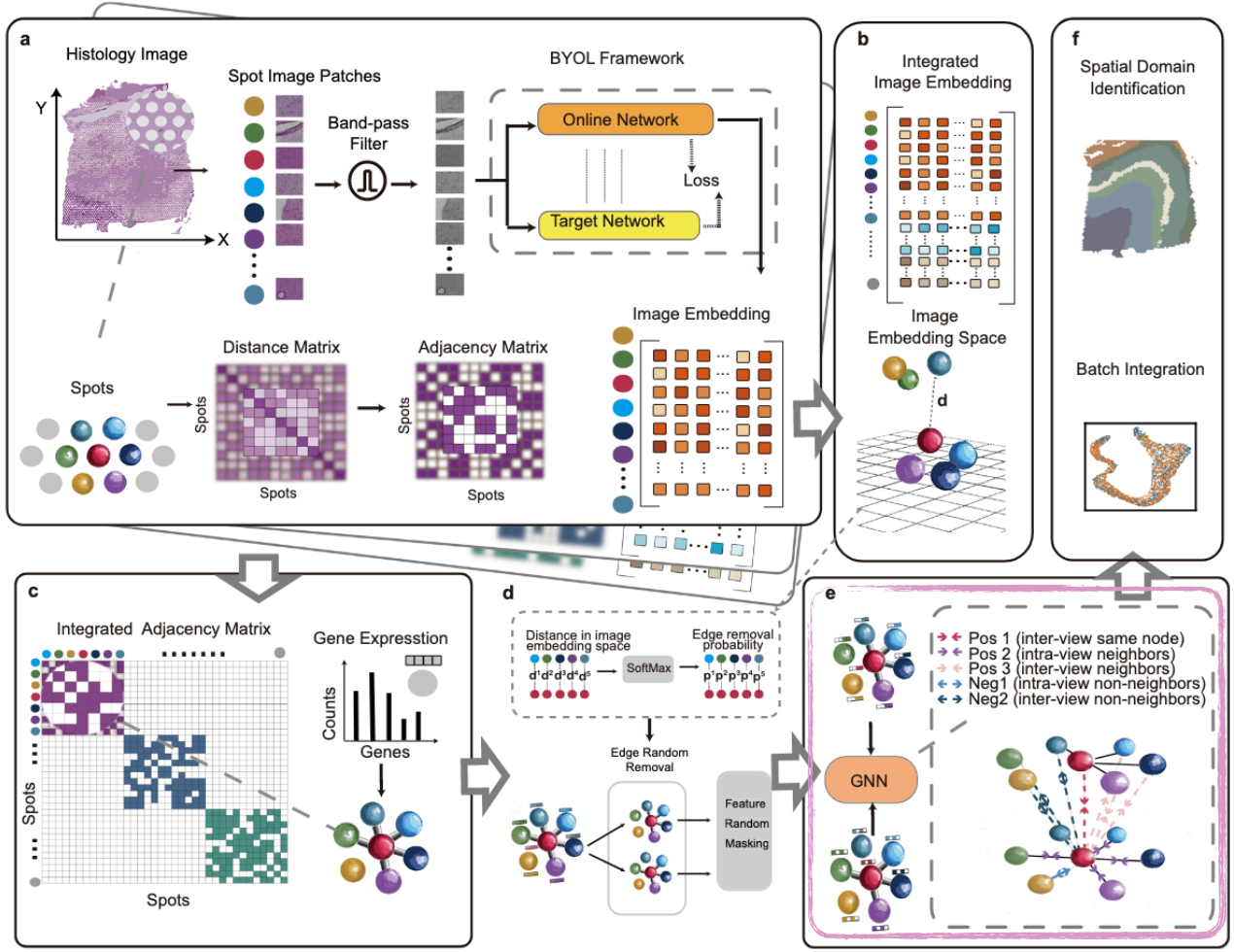
$$H^1 = GNN(G^1, \mathcal{X}^1) \in \mathbb{R}^{TxF'}$$
$$H^2 = GNN(G^2, \mathcal{X}^2) \in \mathbb{R}^{TxF'}$$

- Neighbor contrastive loss:
- **positive samples:**

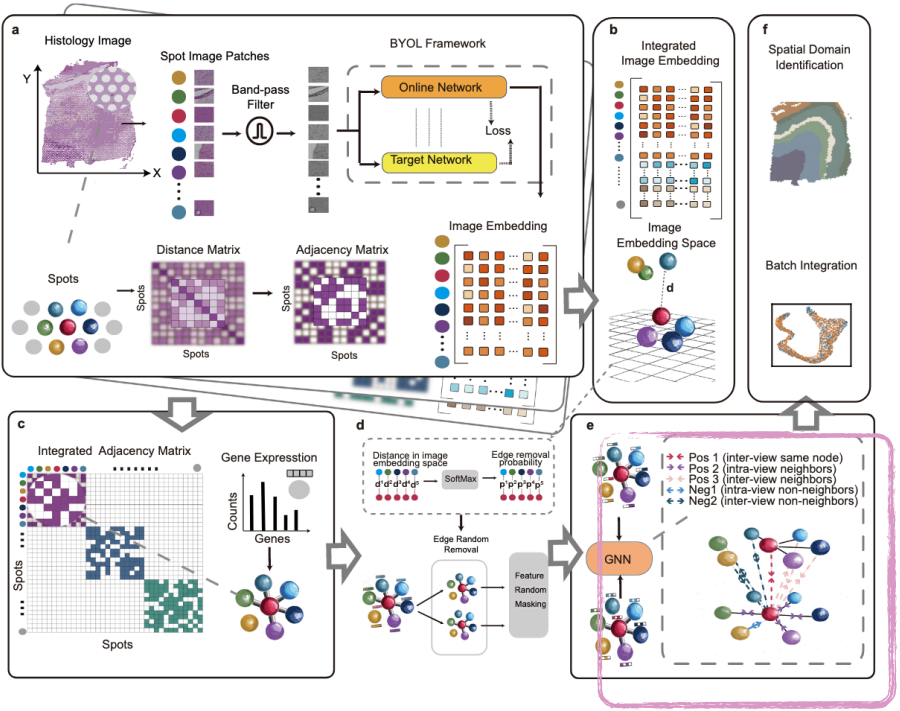
1. embeddings of v_i in G^1 and G^2
2. embeddings of $\mathcal{N}_i$ in G^1
3. embeddings of $\mathcal{N}_i$ in G^1

- **negative samples:**

1. <u>subset</u> of the remaining nodes based on pseudo labels $\{y_i\}_{i=1}^{T}$
2. limited to the same slice



Method overview

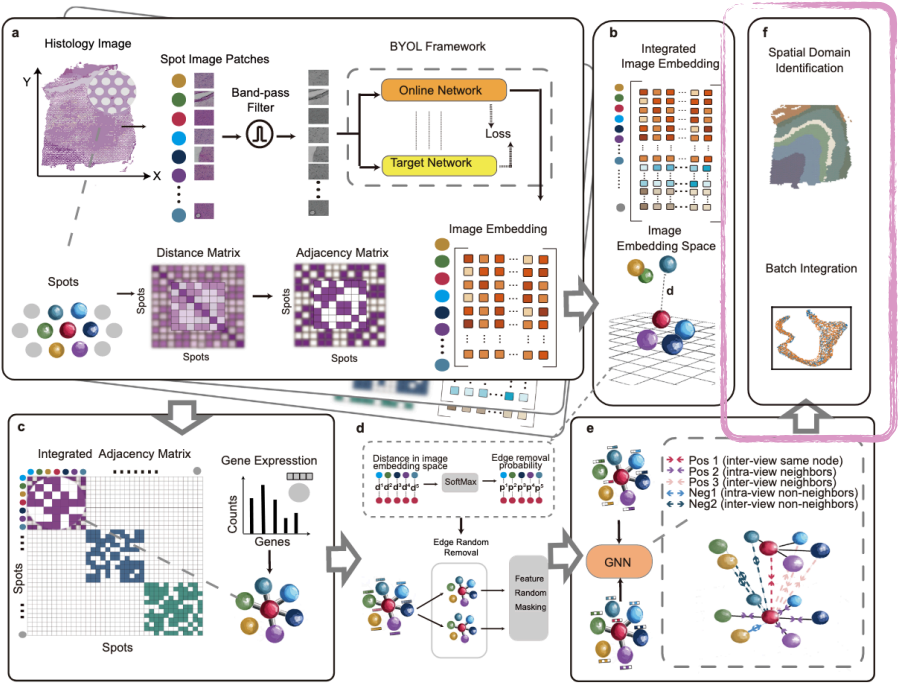


$$\ell(h_i^1) = -\log \frac{\left(f^{1,2}(i,i) + \sum_{v_j \in \mathcal{N}_i^1} f^{1,1}(i,j) + \sum_{v_j \in \mathcal{N}_i^2} f^{1,2}(i,j)\right) \Big/ \mathcal{N}_i}{f^{1,2}(i,i) + \sum_{(j \neq i) \cup (s(v_i) = s(v_j))} \left(f^{1,1}(i,j) + f^{1,2}(i,j)\right)}$$

The final constrastive loss is:

$$l(H^1, H^2) = \frac{1}{2T} \sum_{i=1}^{T} \left[l(h_i^1), l(h_i^2) \right]$$

Method overview



The derived embeddings in both augmented graphs are averaged:

$$H = \frac{1}{2}(H^1 + H^2)$$

The H is clustered using the mclust algorithm. For the number of clusters:

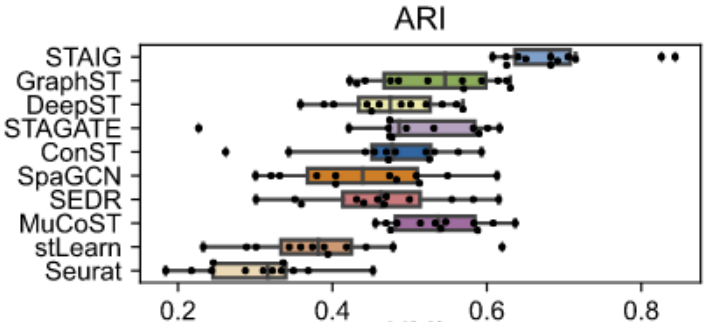
1. If labeled → matches the number of clusters
2. In absence of labels → determined by Silhouette Coefficient (SC)

Datasets

Platform	Tissue Section	#Spots/Bins
10x Visium (Human DLPFC)	151507	4226
	151508	4384
	151509	4789
	151673 (and other)	3639
10x Visium (Zebrafish melanoma)	GSM4838131_Visium_A	2179
	GSM4838132_Visium_B	2677
10x Visium (Human breast)	Human Breast Cancer Section 1	3798
10x Visium (Mouse brain)	Mouse Brain Section 1 (Sagittal Anterior)	2695
	Mouse Brain Section 1 (Sagittal Posterior)	3355
Slide-seqV2	Mouse hippocampus (Puck_200115_08)	52869
	Mouse olfactory bulb (Puck_200127_15)	20139
MERFISH	Mouse visual cortex (Mouse1.AUD_TEA_VIS.242.unexpand)	5995
	Mouse visual cortex (Mouse2.AUD_TEA_VIS.242.unexpand)	2479
	Human middle temporal gyrus (H18.06.006.MTG.4000.expand.rep3)	3970
Stereo-seq	Mouse olfactory bulb	19109
STARmap	Mouse visual cortex	1207

Results

Region Identification in Human Dorsolateral Prefrontal Cortex (DLPFC)

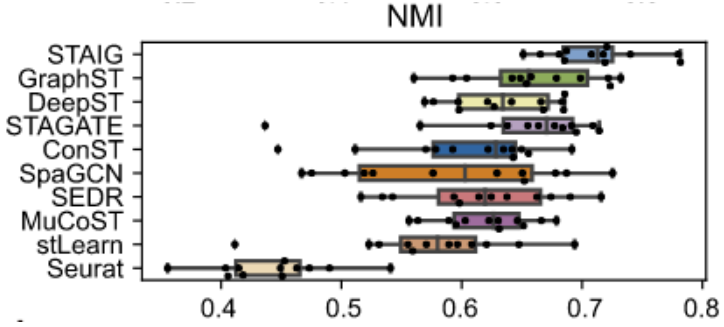


Metrics used:

Adjusted Rand Index (ARI)

Measures clustering agreement while correcting for chance.

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{1 - \mathbb{E}[\text{RI}]}$$



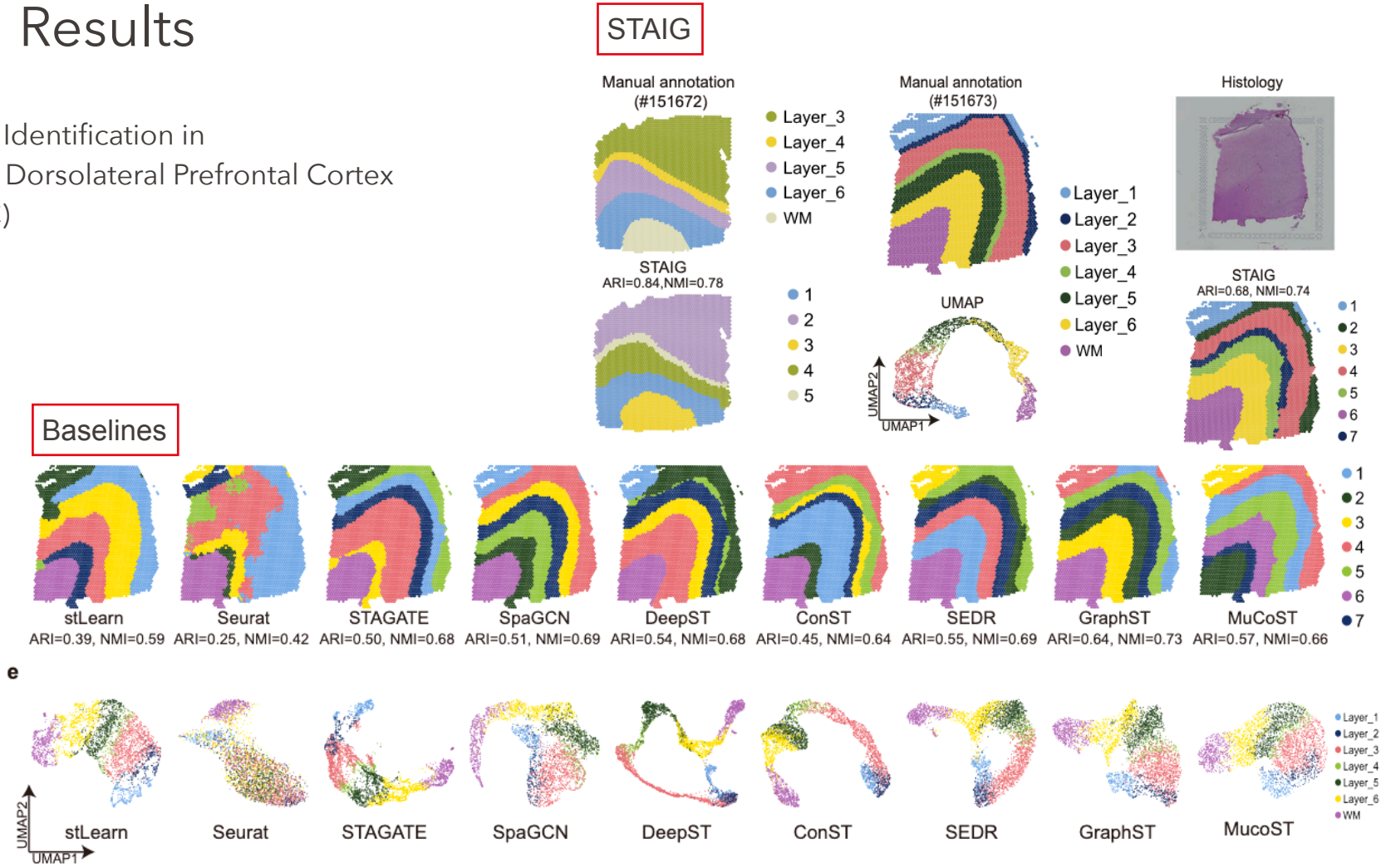
NMI:Normalized Mutual Information (NMI)

Measures shared information between two clusterings, normalized to the sum of their entropies

$$\text{NMI}(U, V) = \frac{2\, I(U, V)}{H(U) + H(V)}$$

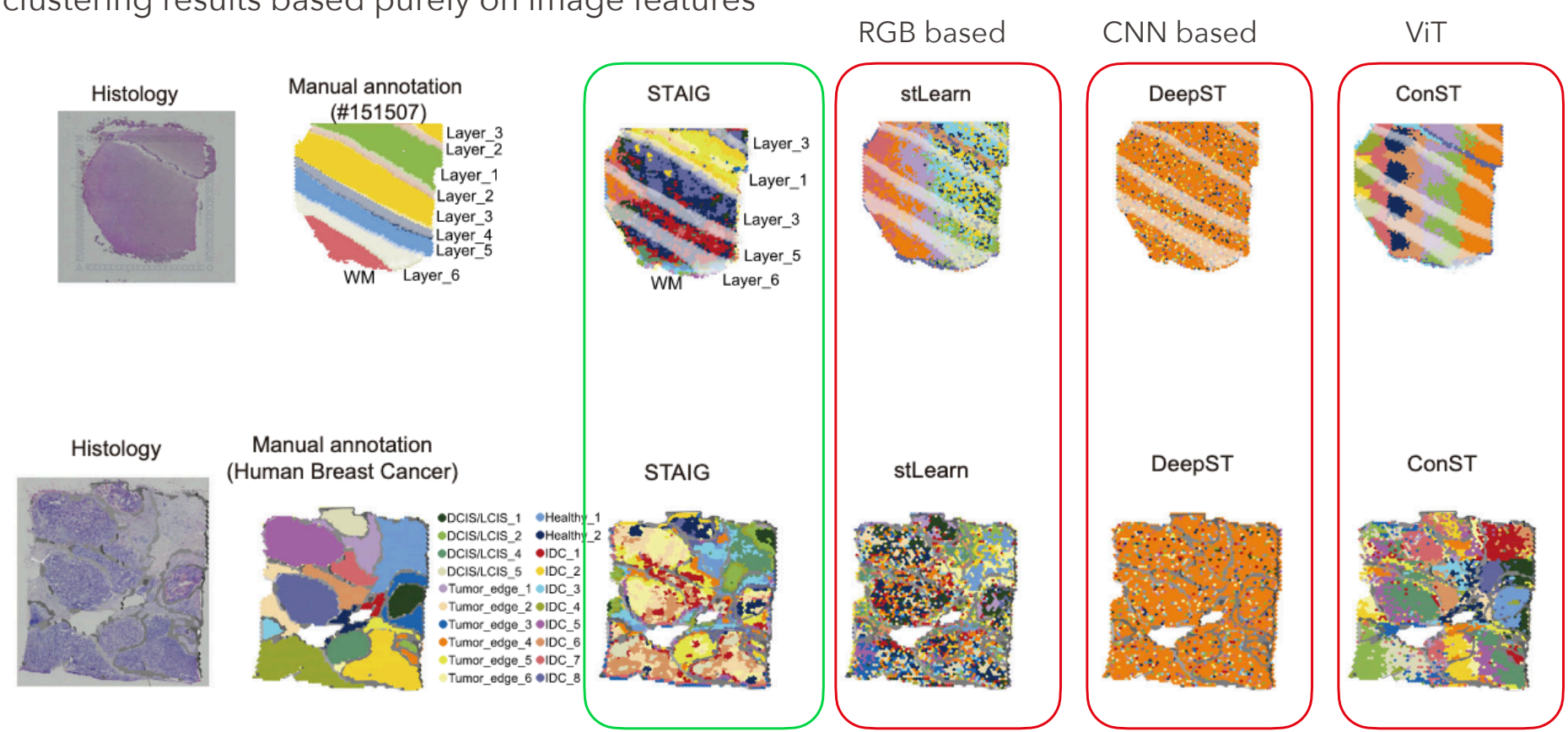
Results

Region Identification in
Human Dorsolateral Prefrontal Cortex
(DLPFC)



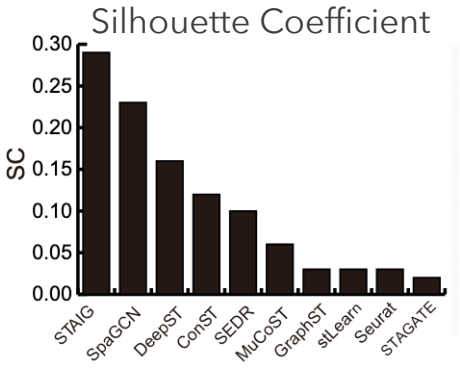
Results

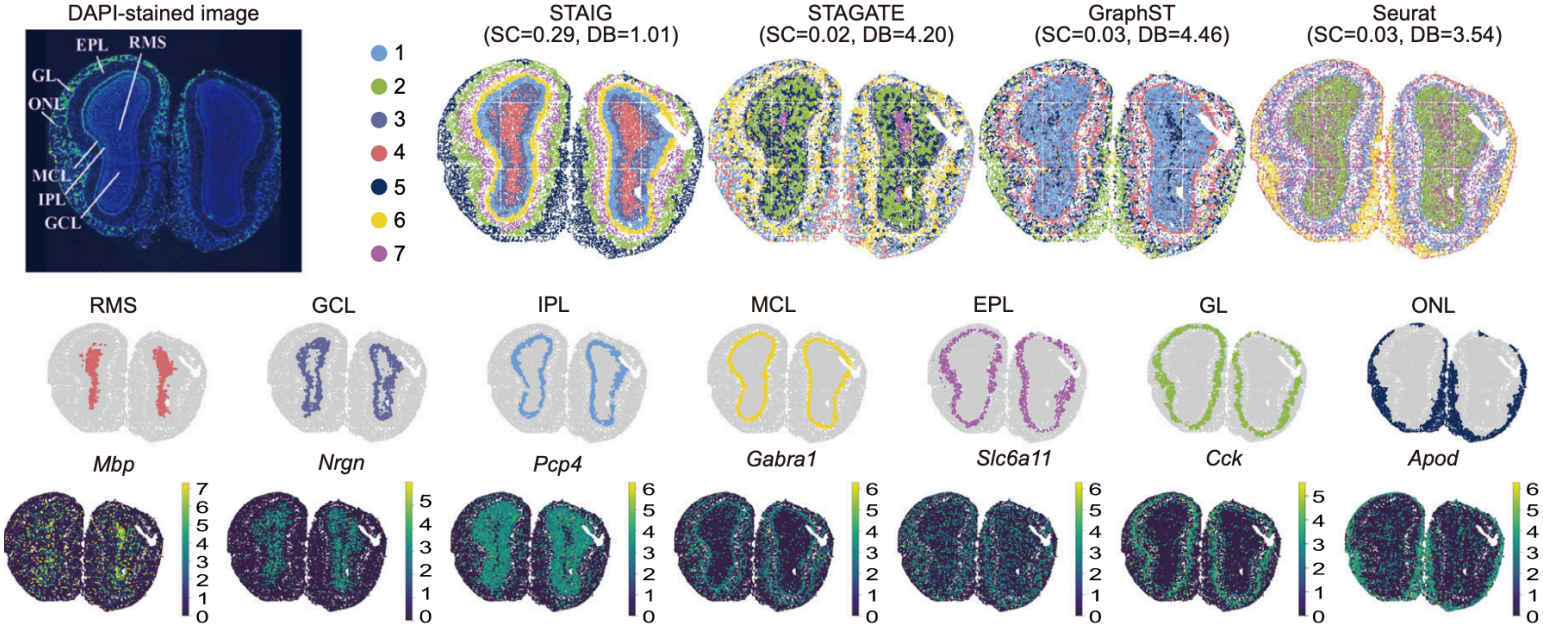
H&E stained image, manual annotations, and KNN clustering results based purely on image features



Results

STAIGs performance when lacking Imaging data
Stereo-seq mouse olfactory bulb tissue





Final Remarks

- Effective multi-modal data integration (elegant image information integration that reduces arbitrary edge biases)
- Alignment-free integration of multiple tissue slices (vertically stacks image embeddings and builds a block-diagonal adjacency matrix)
- Strong empirical performance across different platforms and datasets
- Neighborhood contrastive learning tailored for spatial data

- Graph construction still relies on initial KNN criteria and block-diagonal concatenation.
- Computational complexity (scalability of GNNs and expensive contrastive training)
- Stability issues in G^1, G^2, mitigated with resampling which increased compute & complexity
- Hyper-parameter sensitivity (k in KNN, choice of HVGs, PCA dims, edge-removal probability, neighbor positive and negative set construction)

Leveraging Tumor Heterogeneity: Heterogeneous Graph Representation Learning for Cancer Survival Prediction in Whole Slide Images

Junxian Wu[1,2*] Xinyi Ke[3*] Xiaoming Jiang[4*] Huanwen Wu[3†] Youyong Kong[2,5†]
Lizhi Shao[1†]

- 38th Conference on Neural Information Processing Systems (NeurIPS 2024)

Background

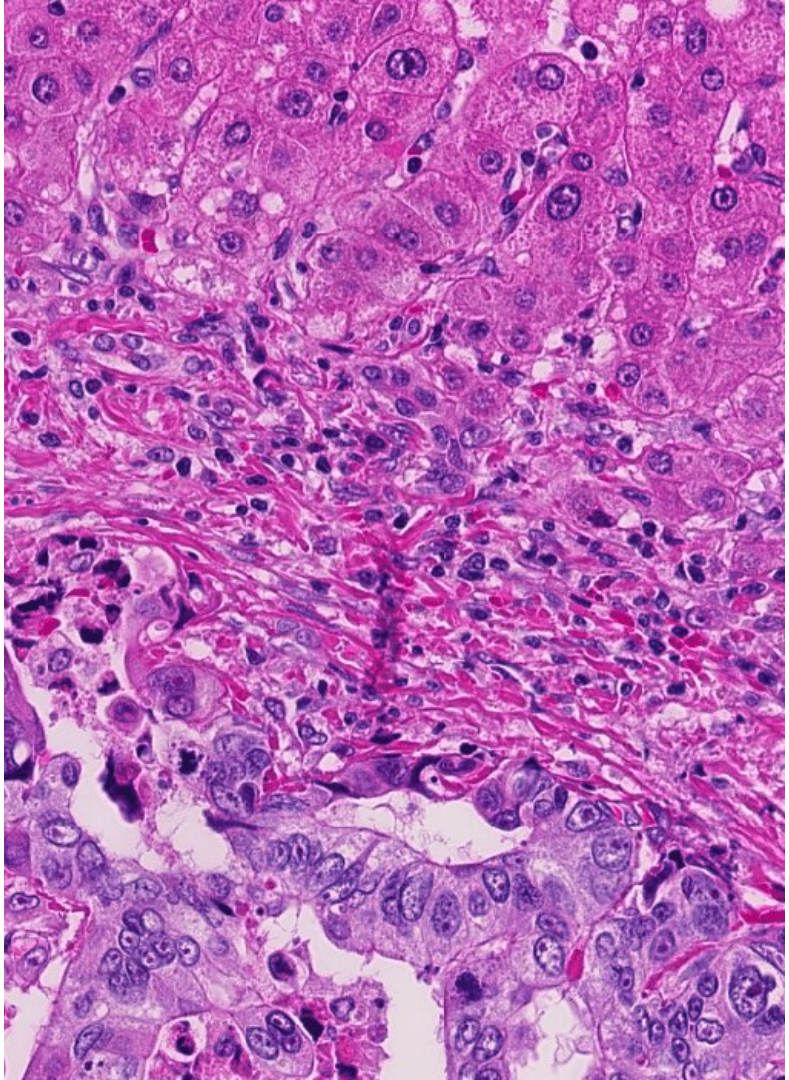
Whole Slide Images (WSI): High resolution digital versions of physical glass microscope slides, enabling viewing, sharing and analysis of histopathology slides without needing the physical slide. The can be created from a slide stained with various methods, including H&E (previous paper…)

- Gold standard in cancer diagnosis
- Rich prognosis information

Survival Prediction: in the context of cancer research is the task of modeling and forecasting the time duration until a specific event occurs (i.e., death of a patient)

- estimated based on the Cox Proportional Hazards (Cox PH) model for survival analysis.

$$\mathscr{L}_{cox} = -\log(L(\theta)) = -\sum_{i:\delta_i=1}\left(\theta_i - \log\sum_{j\in R(t_i)} e^{\theta_j}\right)$$



Motivation

Goal: Cancer Survival Prediction using WSIs.
ProtoSurv: heterogeneous graph model for cancer prognosis prediction

Current approaches & **limitations**:

- The detailed cellular information in WSIs comes at a cost of high memory usage
- To solve this Multiple Instance Learning (MIL) used to obtain slide-level representation
  - Loss of structure information across tissues
  - Struggle to achieve good performance on prognostic prediction tasks

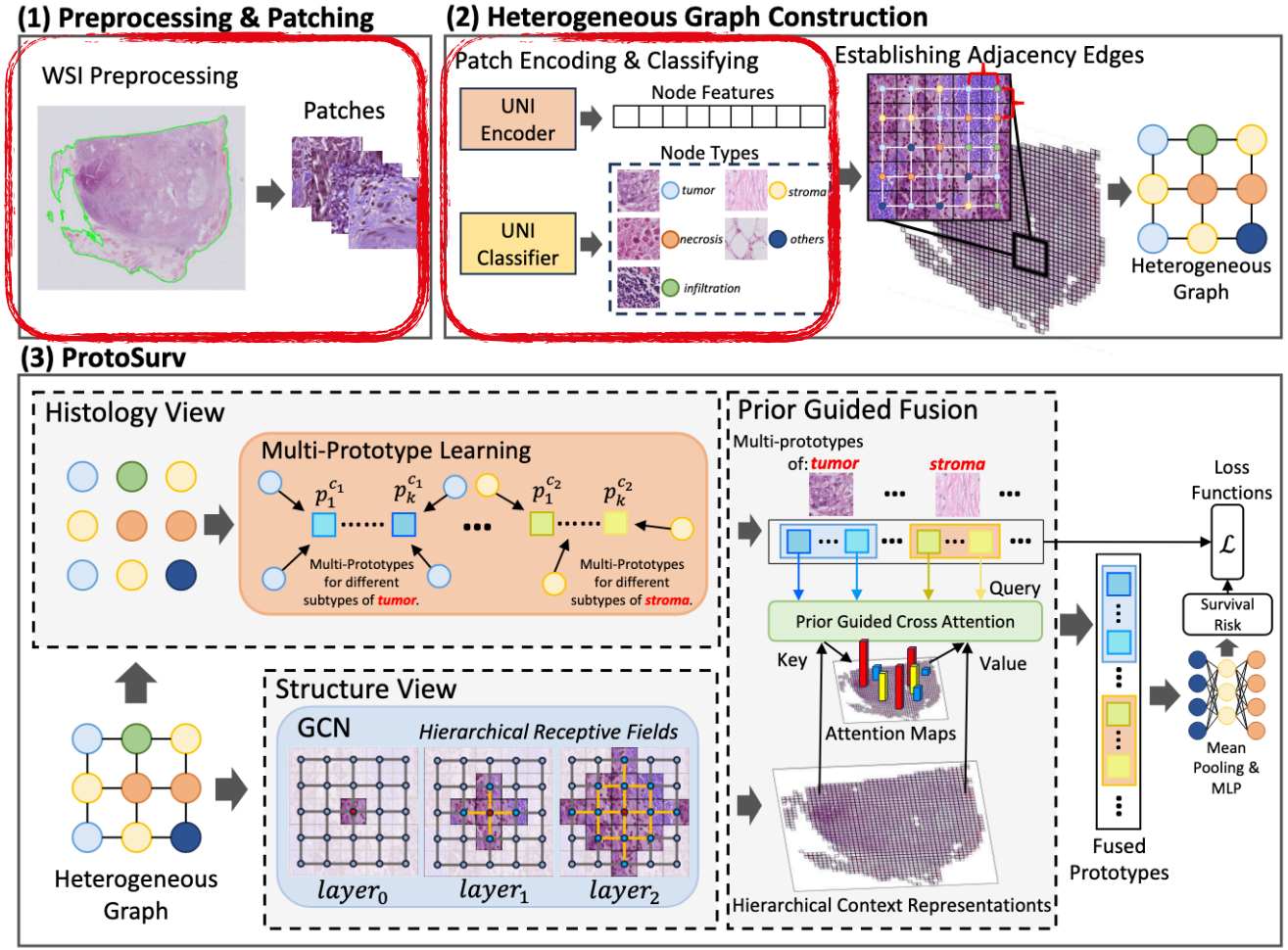
Contributions of **ProtoSurv**:

- Domain knowledge awareness
  - Incorporate prior knowledge of prognostic tissue types
- Extensive validation on five public cancer benchmark datasets and comparison with SOTA

Method overview

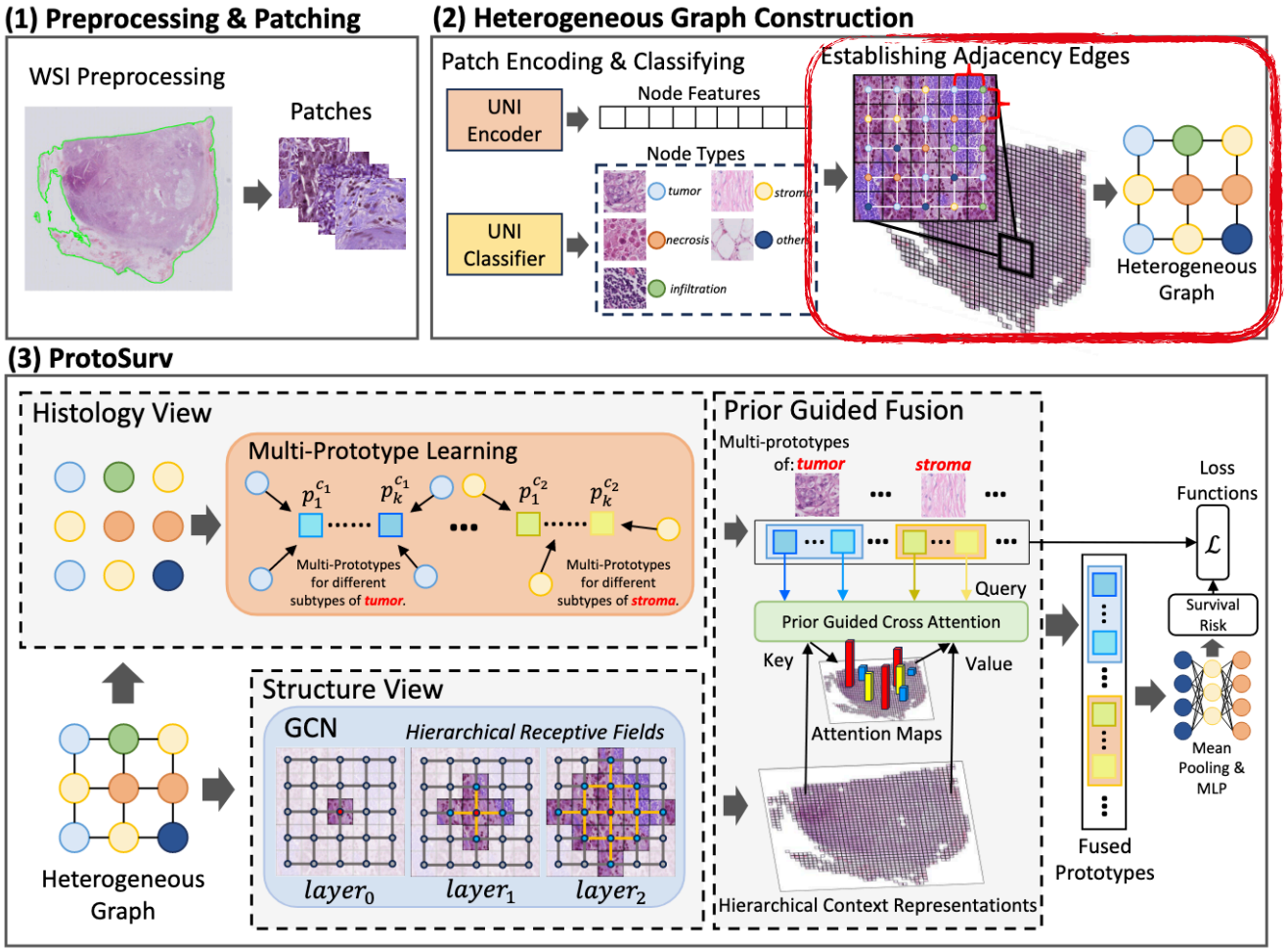
WSIs with 20x magnification are split into non-overlapping **patches** of size 256 x 256

- Use pre-trained **UNI Encoder** to obtain path level embeddings
- Fine-tune **UNI classifier** to obtain the category c_i of each patch ($C = 5$):
  1. Stroma
  2. Tumor
  3. Immune infiltration
  4. Necrosis
  5. others



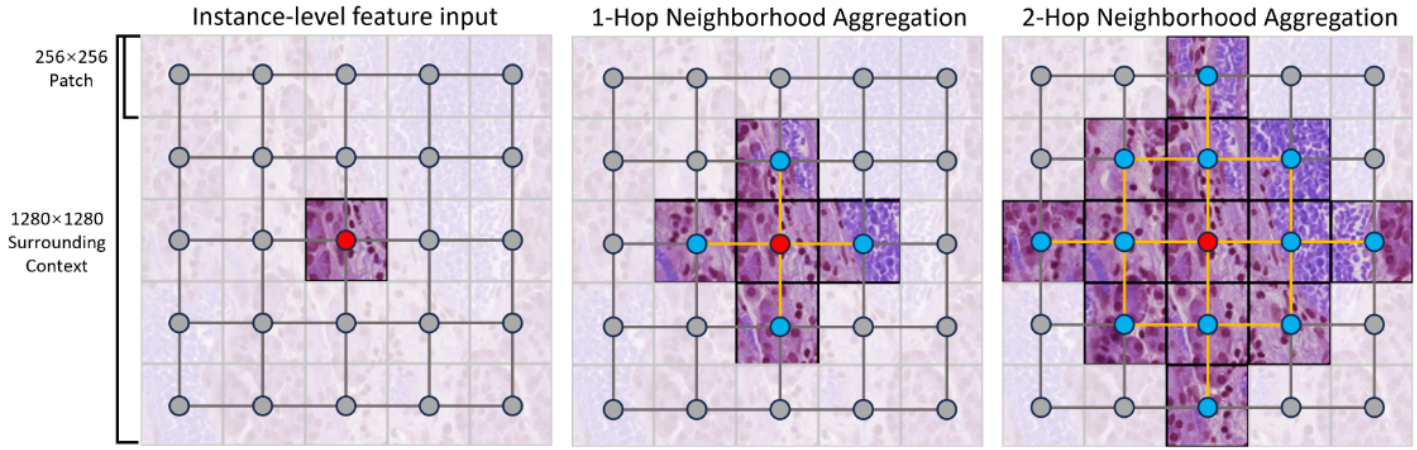
Method overview

Graph construction $G = (V, E)$, where each patch p_i is a node v_i in the graph. Every node v_i within the heterogeneous has a **category** c_i and a $d-$dimensional **feature vector** $x_i \in \mathcal{X}$



Method overview

Structure View (SV): extract multi-hop neighborhood information



- L GCN layers to leverage multi-hop neighborhood information:

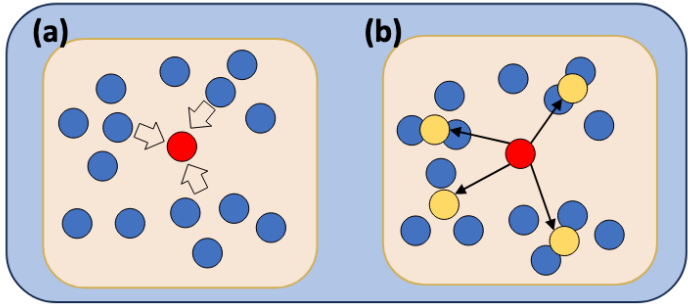
$$h^l = GCN(h^{l-1}, A) \in \mathbb{R}^{Nxd}, \text{ where } h^0 = X$$

$$H' = concat[h^1, h^2, \ldots, h^L] \in \mathbb{R}^{NxLd}$$

$$H = MLP(H') \in \mathbb{R}^{Nxd_h}$$

Method overview

Histology View (HV): learns prototype representations for each category



(a) Calculate the initial prototype p^c_{init} of a certain category using the average function.
(b) Using learnable parameters to shift p^c_{init} to multi-prototypes $P^c_{prior}=\{p^1_c, p^2_c, \ldots, p^k_c\}$, to focus on different clusters (subtypes) within the category.

1. All node features within each category c are averaged to obtain the initial prototype

$$p^c_{init} = MEAN(X_c) \in \mathbb{R}^{1xd}$$

2. K learnable parameters $\{z_1, \ldots, z_k\}$ are used to shift p^c_{init} into multi-prototypes

$$P^c_{prior} = \{p^1_c, p^2_c, \ldots, p^k_c\}$$
$$= \{p^c_{init} + z_1, p^c_{init} + z_2, \ldots, p^c_{init} + z_k\} \in \mathbb{R}^{Kxd}$$

3. cross-attention for updating:

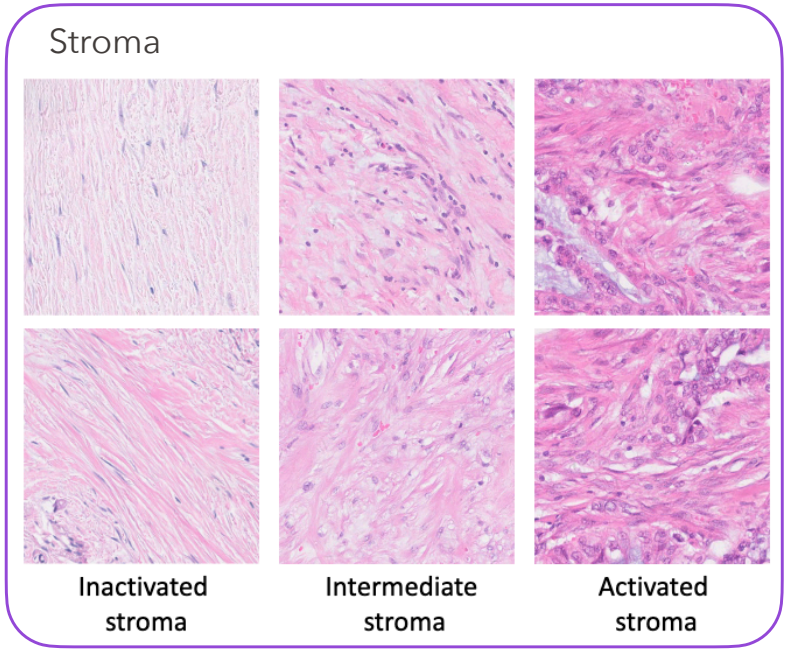
$$P_c = softmax\left(\frac{[P^c_{prior}W^Q][XW^K]^T}{\sqrt{d}}\right)(XW^V) \in \mathbb{R}^{Kxd}$$

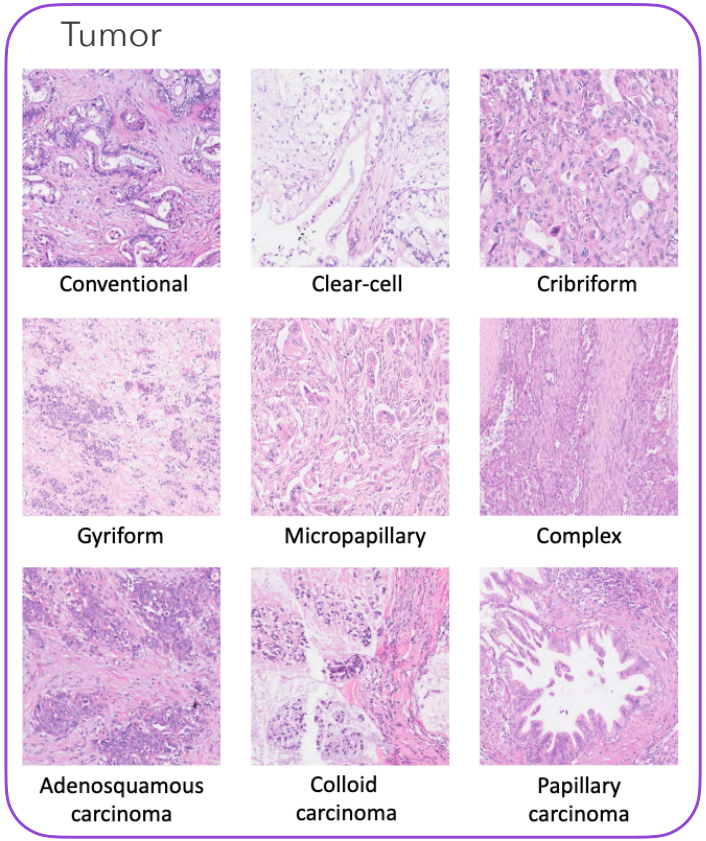
4. Multiple prototypes P for all categories:

$$P = \{P_1, P_2, \ldots, P_c\} \in \mathbb{R}^{CKxd}$$

Method overview

Histology View (HV): illustration of various subtypes of certain tissues, justifying the use of multi-prototype





Method overview

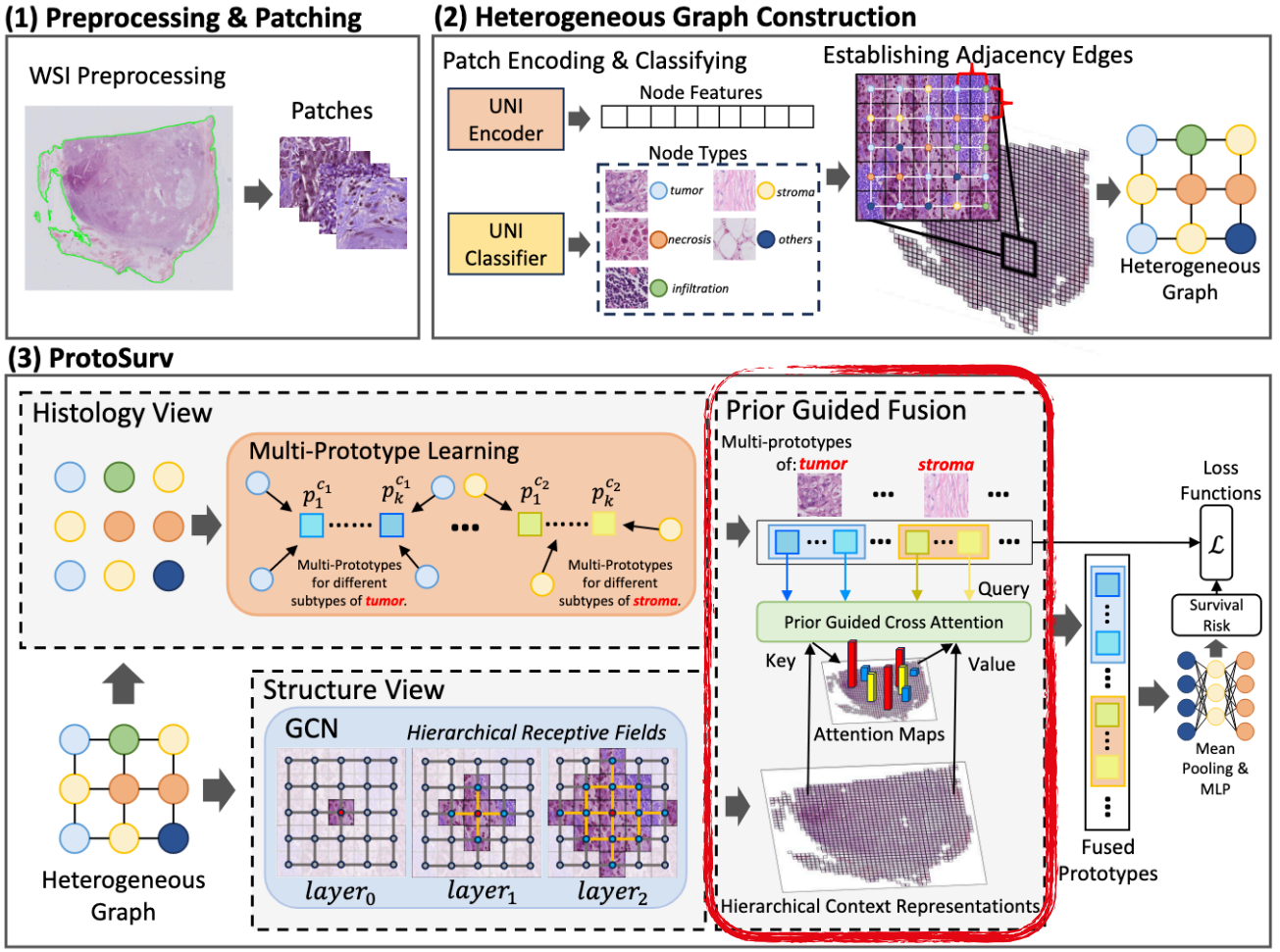
Prior Guided Fusion (PGF)

Cross attention module to aggregate the SV features with the pathological HV multi-prototypes:

$$P_{fusion} = Softmax\left(\frac{[PW^Q][HW^K]^T}{\sqrt{d}}\right)(HW^V) \in \mathbb{R}^{CKxd}$$

$$h_{slide} = MEAN(P) \in \mathbb{R}^{1xd}$$

$Y = g(h_{slide})$, $g(.)$ is an MLP and Y is the predicted survival risk.



Method overview

Overview of the loss function $\mathscr{L}$

$$\mathscr{L} = \mathscr{L}_{cox} + \alpha \mathscr{L}_{comp} + \beta \mathscr{L}_{ortho}$$

Cox regression loss

Compatibility loss

Orthogonality loss

$$\mathscr{L}_{ortho} = \left|\left| \frac{(P^c)^T P^c}{||(P^c)^T P^c||_F} - \frac{I_d}{\sqrt{d}} \right|\right|_F$$

$$s_i^c = MEAN(\gamma(X^c, f(p_k^c)))$$

$$\mathscr{L}_{comp} = \frac{1}{CN} \sum_{i \in N} \left[s_i^{c_i} + \log \sum_{c' \neq c} \exp(-s_i^{c'}) \right]$$

Results

All cancer types of WSIs are from The Cancer Genome Atlas (TCGA) repository

Database	Cases
Breast Invasive Carcinoma (BRCA)	1064
Lower Grade Glioma (LGG)	841
Lung Adenocarcinoma (LUAD)	512
Colon Adenocarcinoma (COAD)	441
Pancreatic Adenocarcinoma (PAAD)	208

Table 1: C-index (mean ± std) over five cancer datasets. The best and second-best results are highlighted in **bold** and underlined.

	PAAD	BRCA	LGG	LUAD	COAD
WSISA	0.573 ± 0.021	0.564 ± 0.054	0.610 ± 0.013	0.576 ± 0.045	0.564 ± 0.034
ABMIL	0.625 ± 0.063	0.657 ± 0.064	0.710 ± 0.048	0.653 ± 0.059	0.647 ± 0.036
TransMIL	0.642 ± 0.037	0.694 ± 0.053	0.739 ± 0.034	0.608 ± 0.040	$\mathbf{0.695 \pm 0.051}$
DeepAttMISL	0.596 ± 0.034	0.634 ± 0.017	0.657 ± 0.076	0.623 ± 0.049	0.638 ± 0.069
Patch-GCN	0.618 ± 0.057	0.647 ± 0.032	0.713 ± 0.054	0.635 ± 0.027	0.652 ± 0.086
DeepGraphConv	0.615 ± 0.032	0.535 ± 0.014	0.617 ± 0.048	0.597 ± 0.037	0.621 ± 0.085
HEAT	0.638 ± 0.030	0.693 ± 0.084	0.741 ± 0.079	0.642 ± 0.031	0.679 ± 0.056
HGSurvNet	0.646 ± 0.064	0.701 ± 0.067	0.746 ± 0.043	0.638 ± 0.064	0.673 ± 0.043
PANTHER	$\mathbf{0.673 \pm 0.082}$	0.699 ± 0.019	0.748 ± 0.046	0.631 ± 0.029	0.635 ± 0.056
ProtoSurv (Ours)	0.669 ± 0.049	$\mathbf{0.720 \pm 0.040}$	$\mathbf{0.774 \pm 0.063}$	$\mathbf{0.658 \pm 0.046}$	0.692 ± 0.045

Results

Ablation study

w/ and w/o the main modules in ProtoSurv

	PAAD	BRCA	LGG	LUAD	COAD
w/o HV	0.618 ± 0.057	0.647 ± 0.032	0.713 ± 0.054	0.635 ± 0.027	0.652 ± 0.086
w/o SV(mean pooling)	0.624 ± 0.032	0.657 ± 0.049	0.706 ± 0.036	0.646 ± 0.051	0.684 ± 0.044
w/o SV(concat pooling)	0.661 ± 0.057	0.713 ± 0.039	0.766 ± 0.044	0.641 ± 0.037	0.688 ± 0.046
w/o PGF(transpose fusion)	0.653 ± 0.024	0.714 ± 0.041	0.724 ± 0.042	0.653 ± 0.055	0.657 ± 0.074
w/o PGF(concat fusion)	0.652 ± 0.040	0.719 ± 0.024	0.712 ± 0.084	0.662 ± 0.064	0.659 ± 0.058
ProtoSurv	0.669 ± 0.049	0.720 ± 0.040	0.774 ± 0.063	0.658 ± 0.046	0.692 ± 0.045

SV = Structure View
HV = Histology View
PGF = Prior Guided Fusion

Effect of tissue category choices

	PAAD	BRCA	LGG	LUAD	COAD
DTC	0.641 ± 0.087	0.684 ± 0.063	0.793 ± 0.056	0.652 ± 0.064	0.681 ± 0.051
CTC	0.656 ± 0.066	0.706 ± 0.058	0.776 ± 0.064	0.658 ± 0.050	0.690 ± 0.041
PTC(proposed)	0.669 ± 0.049	0.720 ± 0.040	0.774 ± 0.063	0.658 ± 0.046	0.692 ± 0.045

DCT = Detailed Tissue Category
CTC = Coarse Tissues Category
PGF = Prior Tissue Category

Effects of compatibility and orthogonality losses

$\mathcal{L}_{comp}$ (α)	$\mathcal{L}_{ortho}$ (β)	PAAD	BRCA	LGG	LUAD	COAD
0	0	0.651 ± 0.057	0.693 ± 0.046	0.765 ± 0.054	0.661 ± 0.052	0.690 ± 0.046
0.1	0.1	0.656 ± 0.049	0.702 ± 0.052	0.769 ± 0.051	0.654 ± 0.046	0.694 ± 0.038
0.1	0.01	0.658 ± 0.055	0.723 ± 0.037	0.773 ± 0.063	0.654 ± 0.041	0.692 ± 0.043
0.01	0.1	0.669 ± 0.049	0.720 ± 0.040	0.774 ± 0.063	0.658 ± 0.046	0.692 ± 0.045
0.01	0.01	0.662 ± 0.043	0.714 ± 0.025	0.770 ± 0.060	0.658 ± 0.049	0.690 ± 0.044

Results

Ablation study

Computational requirements

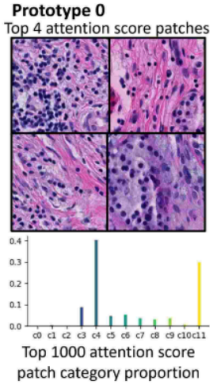
	Time (s)	FLOPs (G)	Model Parameters (M)	Maximum GPU memory usage (MB)
ProtoSurv	0.29	627.3	39.1	5417
ProtoSurv-tiny	0.21	96.5	4.77	1523
PatchGCN	0.12	30.5	1.19	1570

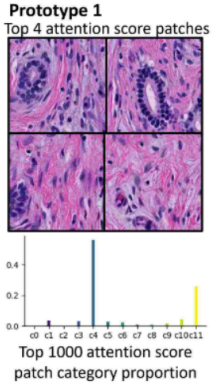
Performance of ProtoSurv-tiny

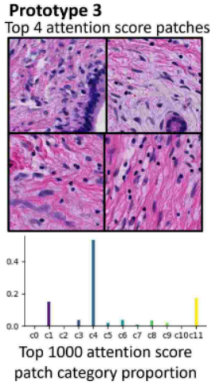
	PAAD	BRCA	LGG	LUAD	COAD
ProtoSurv	0.669 ± 0.049	0.720 ± 0.040	0.774 ± 0.063	0.658 ± 0.046	0.692 ± 0.045
ProtoSurv-tiny	0.687 ± 0.049	0.707 ± 0.044	0.756 ± 0.038	0.664 ± 0.039	0.673 ± 0.039
Patch-GCN	0.618 ± 0.057	0.647 ± 0.032	0.713 ± 0.054	0.635 ± 0.027	0.652 ± 0.086

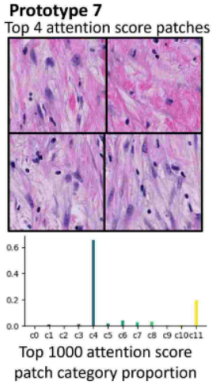
Interpretability

Category 0

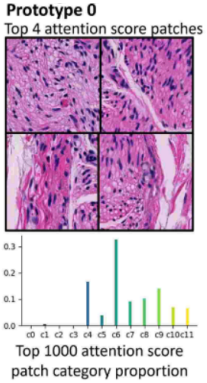


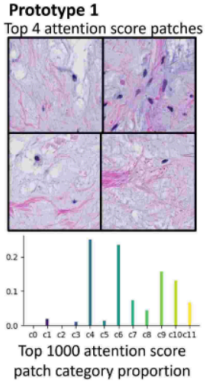


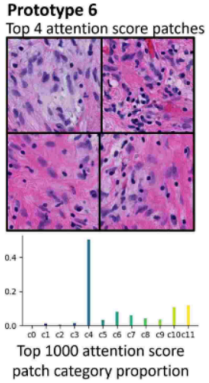


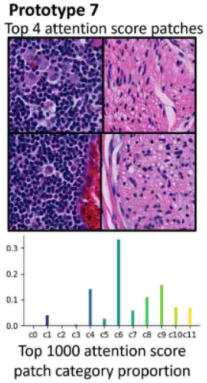


Category 3

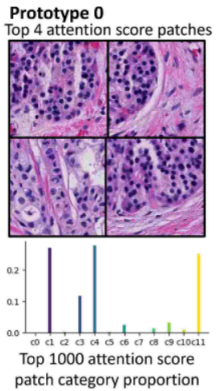


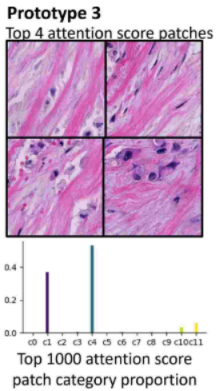


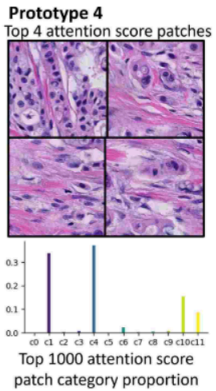


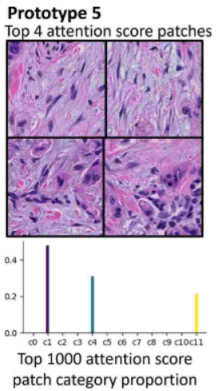


Category 1

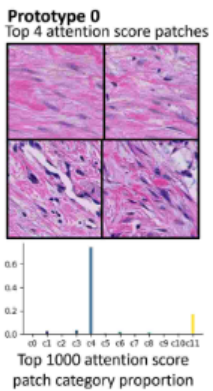


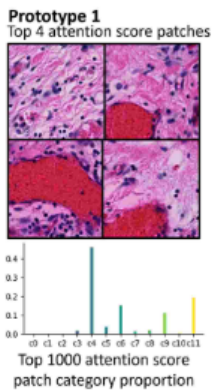


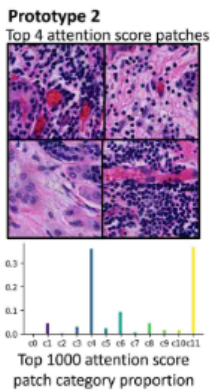


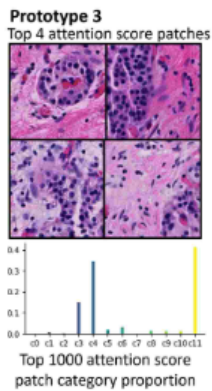


Category 4



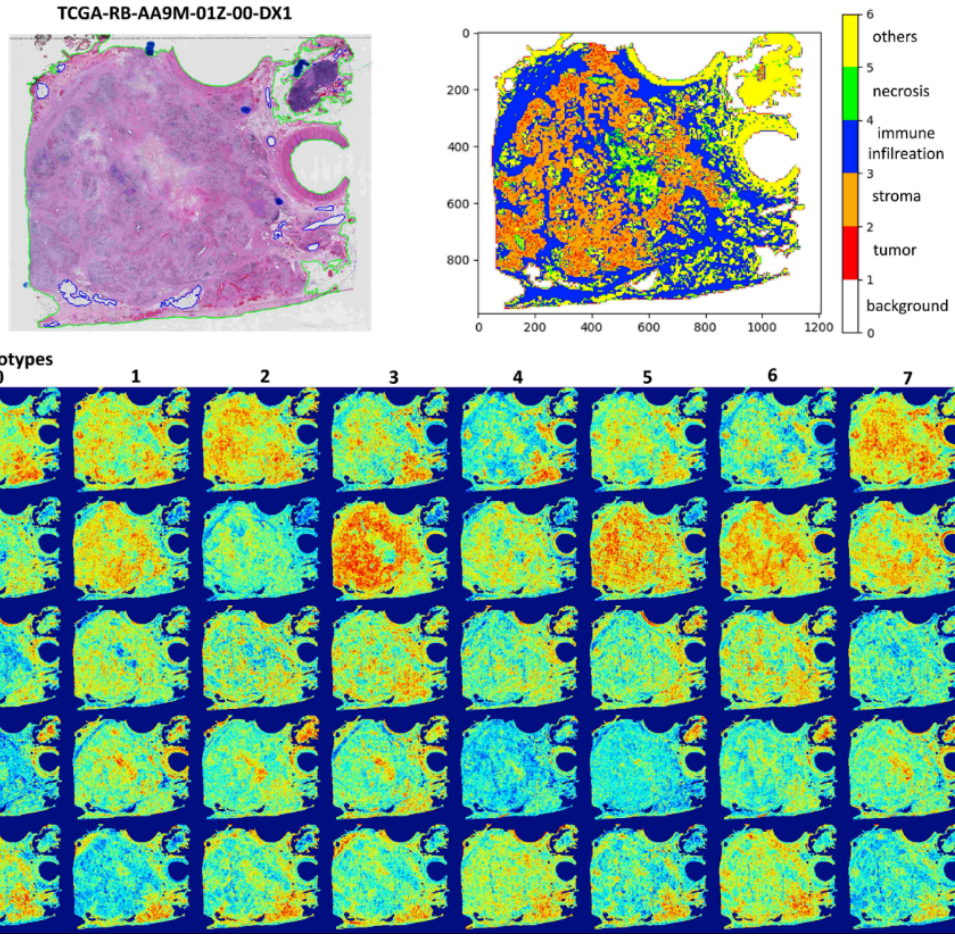






Interpretability

Attention Maps of Each Prototype



Final Remarks

- Incorporating pathological prior knowledge (Histology View)
- Handling intra-tumor heterogeneity (introduces multiple prototypes (K=8) for each tissue category, allowing the model to capture different morphological and prognostic subtypes)
- Robust performance which outperforms or is comparable to the current SOTA models

- Difficulty in obtaining node categories remains an obstacle for its broader application (dependance on additional patch classifier)
- The number of multi-prototypes is fixed and the same for all tissue categories
- Computational complexity (dual-view, cross-attention, GNNs, requires UNI fine-tuning)


Thank you !