Graph contrastive learning of subcellular-resolution spatial transcriptomics improves cell type annotation and reveals critical molecular pathways

Briefings in Bioinformatics, 2025

Qiaolin Lu, Jiayuan Ding, Lingxiao Li, Yi Chang
Reporter: Fengjiao Gong

March 10, 2025

Outline

1. Background
2. Method — Focus
3. Experiments

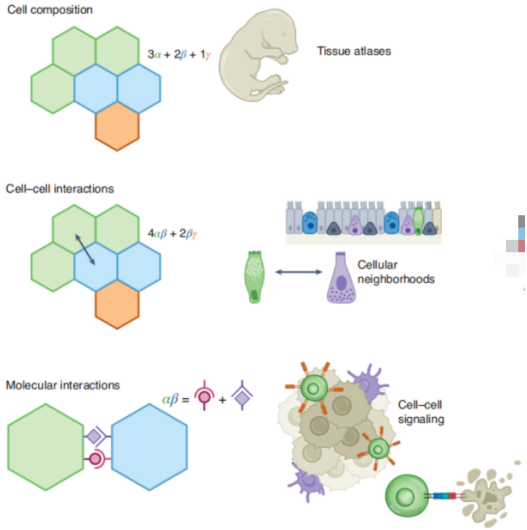
Background

Spatial Transcriptomics(ST) technologies

- enable quantification of RNAs
- within intact tissue sections

Address questions:

- clarify cell-type composition of tissues
- discover cellular spatial interactions rules
- explain molecular interactions between tissue components

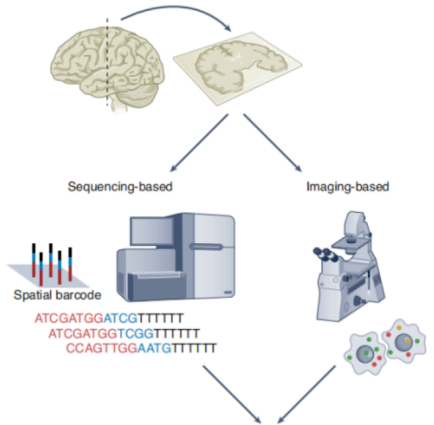


[ref] *The expanding vistas of spatial transcriptomics. Nat Biotechnol 41, 773–782 (2023).*

Background

Spatial Transcriptomics(ST):

- sequencing-based methods(**sST**)
- imaging-based methods(**iST**)
  MERFISH, CosMx SMI, and Xenium



Background

iST

- quantifies gene expression level across cells in space
- directly reveals the subcellular distribution of RNA transcripts
- at the single-molecule resolution

The **subcellular localization** of RNA molecules:

- characterize *cell identity*
- explain *subcellular regulatory mechanisms*

This work — Focus

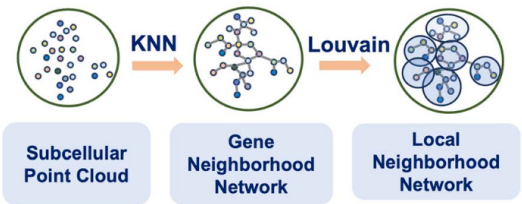
Semi-supervised Graph Contrastive Learning (GCL)-based Algorithm

- learn cell type-specific intracellular spatial distribution of RNA
- first to explicitly model RNA's **subcellular** distribution and community
- improve **cell type annotation** with limited labeled data

Method

- Model
- Unsupervised Contrastive Learning
- Supervised Classification
- Overall Objective

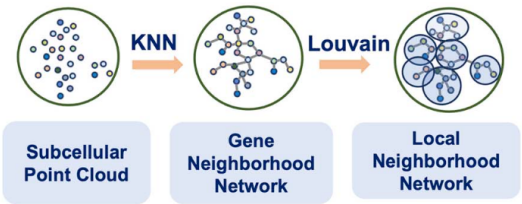
Method — Model



Gene Neighborhood Network G

1. $V = \{v_1, v_2, ..., v_n\}$: the node set with n transcripts in the cell.
2. E: the edge set.
3. $X \in \mathbb{R}^{n \times d}$: node feature.
4. $A \in \mathbb{R}^{n \times n}$: adjacent matrix.

Method — Model

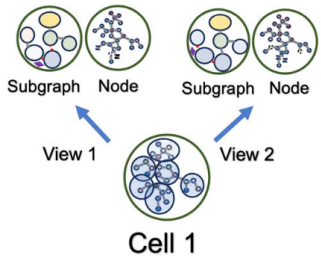


Local Neighborhood Network (graph clustering)

1. Subcellular spatial regions.
2. Subgraphs $\{S_1, S_2, ..., S_k\}$.

Subcellular organization of RNA

Method — Model



Learning graph's **intrinsic** property

- *retain* highly relevant nodes or subgraph
- *eliminate* less critical components

[ref] *Boosting graph contrastive learning via graph contrastive saliency. In: International Conference on Machine Learning, PMLR, The Fortieth International Conference on Machine Learning, 2023.*

Method — Model

Focus uses **multiple GIN layers** to update node embeddings and subgraph embeddings.

$$H_v^{(l+1)} = \text{ReLU}\left(\left((1+\epsilon^{(l)})H_v^{(l)} + \sum_{u \in N(v)} H_u^{(l)}\right) W^{(l)}\right) \tag{1}$$

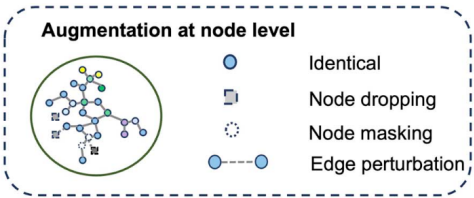
where for the l-th layer

- $H_v^{(l)}$ is the node representation of node v
- $N(v)$ denotes *the set of neighbors* of node v
- $\text{ReLU}(\cdot)$ is the nonlinear activation function
- $\epsilon^{(l)}$ is a learnable parameter
- $W^{(l)}$ is the weight matrix

Method — Unsupervised Contrastive learning

1. Augmentation
   - Node Augmentation
   - Subgraph Augmentation
   - Sampling Strategy
2. Loss Design

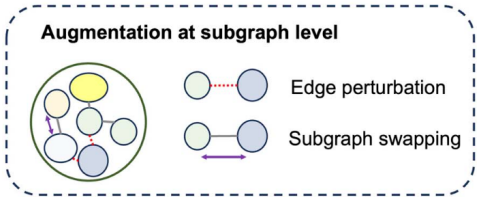
Method — Unsupervised Contrastive learning



Node Augmentation:

- Identical — preserve the original state.
- Node dropping — remove nodes.
- Node masking — conceal node attributes.
- Edge perturbation — delete edges connecting nodes.

Method — Unsupervised Contrastive learning



Subgraph Augmentation:

- Edge perturbation — remove edges between subgraphs.
- Subgraph swapping — exchange of two subgraphs.

Method — Unsupervised Contrastive learning

Sampling Strategy

Gumbel-Softmax strategy

- sample from a discrete distribution in a differentiable way
- calculate Gumbel-Softmax probability as node importance score

[ref] *Categorical reparameterization with Gumbel-Softmax. 5th International Conference on Learning Representations,2017.*

Method — Unsupervised Contrastive learning

Sampling Strategy

For node v with augmentation i:

1. Node dropping & masking:

$$p_{v_i} = 1 - \text{GumbelSoftmax}\left(\text{MLP}_i\left(H_v\right)\right) \tag{2}$$

$$p_v = \sum_i p_{v_i} \qquad \textit{Node Score} \tag{3}$$

2. Edge perturbation:

$$\text{Average}\{\text{related } p_v\}$$

Method — Subgraph Augmentation

For subgraph S with m nodes:

1. Subgraph swapping:

$$p_s = \frac{1}{m} \sum_{i=1}^{m} p_v \qquad \textit{Subgraph Score} \tag{4}$$

2. Edge perturbation:

$$\text{Average}\{\text{related } p_v\}$$

Method — Unsupervised Contrastive learning

Augmented **graph embedding** from ResGCN:

$$H_G = \sum_{v=1}^{n'} H_v^{(t)}, \tag{5}$$

where

- n' — the number of nodes.
- $H_v^{(t)}$ — the node embedding at the t-th layer.

Method — Unsupervised Contrastive Learning

Loss Design

Graph Contrastive Learning (GCL):

- **Positive pair**
  Two augmented views originating from the same input graph.
- **Negative pair**
  Two views derived from distinct input graphs.

Method — Unsupervised Contrastive Learning

Contrastive learning loss for cell:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{H}_{G_i}, \mathbf{H}_{G_j})/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(\mathbf{H}_{G_i}, \mathbf{H}_{G_k})/\tau)} \tag{6}$$

where

- (i, j) represents a positive pair of samples from the same cell
- (i, k) denotes a randomly sampled pair from the batch
- $2N$ is the number of generated graph views
- τ is the temperature parameter

Method — Unsupervised Contrastive Learning

Contrastive learning loss for batch:

$$L_{cl} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)] \tag{7}$$

Method — Supervised Classification

Supervised classification loss:

$$L_{cls} = -\sum_{k=1}^{N} y \log(\hat{y}) \tag{7}$$

where y and $\hat{y}$ denote ground truth labels and predicted labels, respectively.

Method — Overall Objective

Overall objective of Focus:

$$L = L_{cl} + \lambda L_{cls} \tag{8}$$

where λ is a balance hyper-parameter.

Method — Focus

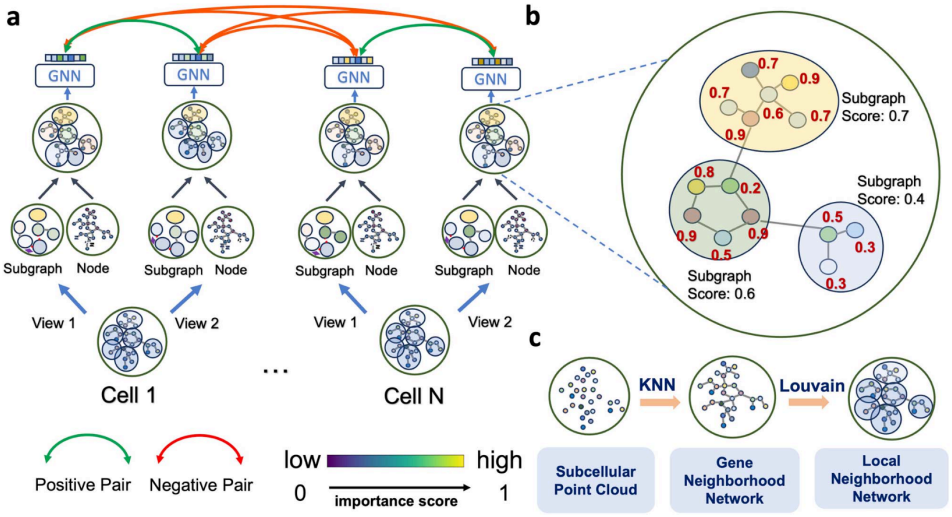


Figure 1: An overview of the Focus framework.

Experiments

Table 1: Summary of Datasets.

Dataset	CosMx Lung	CosMx Kidney	MERFISH MOp	Xenium DCIS
Cells	766313	500000	300000	300000
Genes	960	960	258	313
Types	18	8	24	19
Tissues	Human Lung	Human Kidney	Mouse Cortex	Human Breast

Experiments

1. Cell Type Annotation Across Diverse iST Platforms
2. Ablation Studies on Augmentation Strategies
3. Enriched Cell Type-specific Pathways from Graph-based Gene Importance Scores

Cell Type Annotation Across Diverse iST Platforms

Table 2: Performance of different methods when the reference and query samples come from the same patient or mouse.

Dataset	CosMx Lung		CosMx Kidney		MERFISH MOp		Xenium DCIS	
Model	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Focus	$\underline{0.904 \pm 0.009}$	$\underline{0.704 \pm 0.027}$	$\mathbf{0.688 \pm 0.013}$	$\mathbf{0.694 \pm 0.006}$	$\mathbf{0.948 \pm 0.023}$	$\mathbf{0.909 \pm 0.082}$	$\mathbf{0.871 \pm 0.006}$	$\mathbf{0.712 \pm 0.030}$
scDeepSort	0.725 ± 0.026	0.391 ± 0.012	0.377 ± 0.005	0.248 ± 0.007	0.868 ± 0.052	0.685 ± 0.025	N/A	N/A
CellTypist	0.634 ± 0.018	0.429 ± 0.013	0.478 ± 0.028	0.495 ± 0.010	N/A	N/A	0.371 ± 0.011	0.144 ± 0.002
TOSICA	0.854 ± 0.021	0.638 ± 0.001	$\underline{0.668 \pm 0.002}$	$\underline{0.664 \pm 0.002}$	$\underline{0.942 \pm 0.004}$	0.885 ± 0.059	0.381 ± 0.044	0.176 ± 0.005
ACTINN	0.852 ± 0.002	0.566 ± 0.029	0.569 ± 0.009	0.551 ± 0.022	0.938 ± 0.011	0.853 ± 0.059	$\underline{0.593 \pm 0.013}$	0.193 ± 0.004
Tacco	$\mathbf{0.916 \pm 0.012}$	$\mathbf{0.729 \pm 0.004}$	0.530 ± 0.030	0.522 ± 0.016	0.939 ± 0.011	$\underline{0.902 \pm 0.047}$	0.525 ± 0.001	$\underline{0.207 \pm 0.007}$
scDot	0.844 ± 0.009	0.637 ± 0.029	0.481 ± 0.028	0.506 ± 0.033	0.922 ± 0.029	0.868 ± 0.080	N/A	N/A

Bold text indicates the best, underlined text indicates the second-best, and 'N/A' means no meaningful results.

Cell Type Annotation Across Diverse iST Platforms

Table 3: Performance of different methods when the reference and query samples come from the different patients or mice.

Dataset	CosMx Lung		CosMx Kidney		MERFISH MOp	
Model	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
Focus	**0.692 ± 0.119**	**0.399 ± 0.173**	**0.715 ± 0.025**	**0.708 ± 0.020**	**0.934 ± 0.010**	**0.865 ± 0.048**
scDeepSort	0.462 ± 0.059	0.150 ± 0.045	0.173 ± 0.119	0.065 ± 0.062	0.854 ± 0.033	0.680 ± 0.071
CellTypist	0.450 ± 0.112	0.214 ± 0.018	0.450 ± 0.066	0.472 ± 0.062	N/A	N/A
TOSICA	0.634 ± 0.202	0.386 ± 0.158	0.709 ± 0.002	0.703 ± 0.017	0.925 ± 0.006	0.825 ± 0.006
ACTINN	0.636 ± 0.047	0.299 ± 0.057	0.584 ± 0.009	0.543 ± 0.023	0.931 ± 0.008	0.832 ± 0.039
Tacco	0.640 ± 0.080	0.363 ± 0.146	0.575 ± 0.049	0.561 ± 0.056	0.922 ± 0.011	0.850 ± 0.034
scDot	0.663 ± 0.134	0.390 ± 0.174	0.584 ± 0.062	0.554 ± 0.069	0.918 ± 0.001	0.847 ± 0.047

Bold text indicates the best, underlined text indicates the second-best, and 'N/A' means no meaningful results.

Ablation Studies on Augmentation Strategies

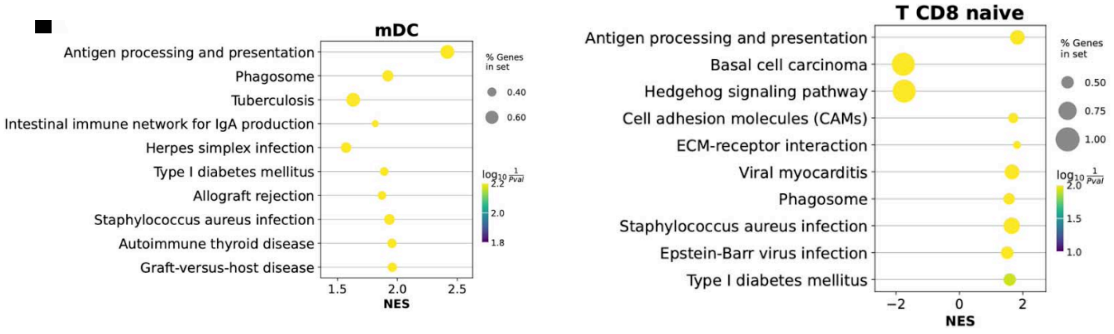
Table 4: Ablation studies of different data augmentation strategies.

Augmentation Level	Augmentation Type*	CosMx Lung		Xenium DCIS		MERFISH MOp	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Node	Identical	0.852 ± 0.044	0.663 ± 0.023	0.838 ± 0.012	0.701 ± 0.021	0.947 ± 0.032	0.911 ± 0.021
Node	ND	0.872 ± 0.021	0.710 ± 0.017	0.842 ± 0.019	0.729 ± 0.021	0.970 ± 0.015	0.961 ± 0.018
Node	NM	0.863 ± 0.018	0.698 ± 0.018	0.832 ± 0.020	0.711 ± 0.019	0.954 ± 0.009	0.922 ± 0.010
Node	EP	0.869 ± 0.011	0.686 ± 0.014	0.834 ± 0.007	0.707 ± 0.006	0.967 ± 0.021	0.944 ± 0.011
Node	ND & NM	0.882 ± 0.015	0.699 ± 0.012	0.855 ± 0.011	0.731 ± 0.023	0.944 ± 0.014	0.887 ± 0.025
Node	ND, NM & EP	**0.901 ± 0.013**	0.691 ± 0.021	**0.871 ± 0.018**	0.733 ± 0.024	0.965 ± 0.017	0.921 ± 0.023
Subgraph	EP	0.857 ± 0.029	0.671 ± 0.030	0.852 ± 0.032	0.711 ± 0.021	0.937 ± 0.033	0.889 ± 0.029
Subgraph	SW	0.873 ± 0.016	0.681 ± 0.031	0.858 ± 0.028	0.732 ± 0.022	0.945 ± 0.018	0.889 ± 0.019
Subgraph	EP & SW	0.862 ± 0.022	0.672 ± 0.028	0.856 ± 0.026	0.700 ± 0.018	0.941 ± 0.026	0.885 ± 0.024
Node & Subgraph	ALL	0.901 ± 0.021	**0.712 ± 0.008**	0.867 ± 0.019	**0.734 ± 0.021**	**0.975 ± 0.011**	**0.967 ± 0.018**

*Identical: no augmentation; ND: node dropping; NM: node masking; EP: edge perturbation; SW: subgraph swapping; ALL: all augmentations including ND, NM & EP from node level and EP & SW from subgraph level. Bold text indicates the best and underlined text indicates the second-best.

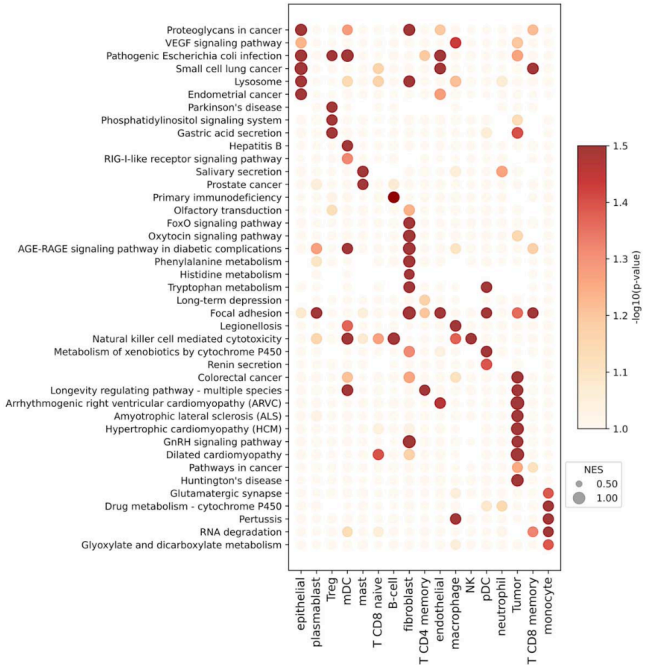
Enriched Cell Type-specific Pathways from Graph-based Gene Importance Scores

CosMx Lung dataset — cellType: mDC + T CD 8 naive



[ref] *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences, 2005*

Heatmap


Proteoglycans in cancer
VEGF signaling pathway
Pathogenic Escherichia coli infection
Small cell lung cancer
Lysosome
Endometrial cancer
Parkinson's disease
Phosphatidylinositol signaling system
Gastric acid secretion
Hepatitis B
RIG-I-like receptor signaling pathway
Salivary secretion
Prostate cancer
Primary immunodeficiency
Olfactory transduction
FoxO signaling pathway
Oxytocin signaling pathway
AGE-RAGE signaling pathway in diabetic complications
Phenylalanine metabolism
Histidine metabolism
Tryptophan metabolism
Long-term depression
Focal adhesion
Legionellosis
Natural killer cell mediated cytotoxicity
Metabolism of xenobiotics by cytochrome P450
Renin secretion
Colorectal cancer
Longevity regulating pathway - multiple species
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Amyotrophic lateral sclerosis (ALS)
Hypertrophic cardiomyopathy (HCM)
GnRH signaling pathway
Dilated cardiomyopathy
Pathways in cancer
Huntington's disease
Glutamatergic synapse
Drug metabolism - cytochrome P450
Pertussis
RNA degradation
Glyoxylate and dicarboxylate metabolism
epithelial
plasmablast
Treg
mDC
mast
T CD8 naive
B-cell
fibroblast
T CD4 memory
endothelial
macrophage
NK
pDC
neutrophil
Tumor
T CD8 memory
monocyte
-log10(pvalue)
1.5
1.4
1.3
1.2
1.1
1.0
NES
0.50
1.00

Conclusion

Focus

- Enhance *cell type annotation* by leveraging transcript's subcellular and spatial community information.
- To validate its *generalizability and robustness* on a broader range of cell types and tissues.

Thanks!