Diffusion Generative Modeling for Spatially Resolved Gene Expression Inference from Histology Images

Sichen Zhu*[1], Yuchen Zhu*[2], Molei Tao†[2], Peng Qiu†[1]

[1]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA
[2]School of Mathematics, Georgia Institute of Technology, GA, USA

Georgia Tech
Machine Learning Center

Introduction

Spatial Transcriptomics (ST) allows a high-resolution measurement of RNA sequence abundance by systematically connecting cell morphology depicted in Hematoxylin and Eosin (H&E) stained histology images to spatially resolved gene expressions.

- ST is a time-consuming, expensive yet powerful experimental technique that provides new opportunities to understand cancer mechanisms at a fine-grained molecular level.
- H&E stained images are enriched in clinical settings due to their low cost and wide application.

Can we develop a machine learning tool to computationally infer spatially resolved gene expression solely based on histology images?
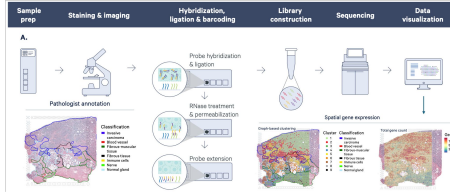
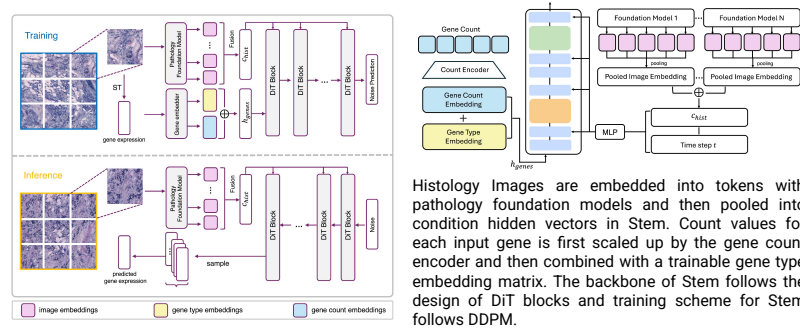## Background

Computational Pathology Foundation Models:

★ CONCH (Lu et al., 2024):a vision language foundation model for histopathology (Pretrained vision-language encoders, vision encoder: ViT-B/16, 90M params; text encoder: L12-E768-H12, 110M params)
★ UNI (Chen et al., 2024): Pretrained ViT-L/16 via DINOv2 for multi-purpose evaluation on histopathology images
★ Virchow-2, H-Optimus-0, etc…

Diffusion Model for Multimodal data and Conditional Generation:

Diffusion Model - a probabilistic model that is designed to learn data distribution. A diffusion model consists of two stochastic processes: forward process perturbs the data distribution to a simple distribution that could be easily sample from; backward process iteratively denoises the noisy data back to the original data distribution.
Diffusion Conditional Generation - modeling the conditional distribution of original data points. Examples for conditions: class label, text, or histology images in our settings.

Diffusion Generative Modeling of Gene Expressions

Histology Images are embedded into tokens with pathology foundation models and then pooled into condition hidden vectors in Stem. Count values for each input gene is first scaled up by the gene count encoder and then combined with a trainable gene type embedding matrix. The backbone of Stem follows the design of DiT blocks and training scheme for Stem follows DDPM.

Results

HER2ST dataset: 36 breast tissue sections from 8 individuals.
Gene selection: high mean high variance gene (HMHVG) or differentially expressed genes (DEG)
Hold out one slide for testing and trained on the other slides.
Evaluation (calculated in log2 space):
❑ Pearson Correlation Coefficient (PCC)
❑ Mean absolute error (MAE)
❑ Mean square error (MSE)
❑ Relative variance distance (RVD)

Pathologist's annotation

3: immune: APC, B−cell, T−cell
1: immune: B/plasma cells
2: adipose
6: cancer : connective tissue
0: cancer 1
4: cancer in situ : immune rich
5: cancer 2
background

Gene-expression-based clustering via nearest neighbor approach
Image Source: (Alma et al. Nature Communications, 2021)

Accurate predictions for selected genes:

Model	HMHVG						DEG					
	PCC-10↑	PCC-50↑	PCC-300↑	MAE↓	MSE↓	RVD↓	PCC-10↑	PCC-50↑	PCC-300↑	MAE↓	MSE↓	RVD↓
HisToGene	0.6812	0.6345	0.5250	0.9367	1.3468	10.3407	0.6816	0.6369	0.5112	0.8791	1.2627	9.7057
BLEEP	0.7727	0.7141	0.5652	0.8328	1.2428	0.6025	0.7711	0.7188	0.5518	0.7590	1.1297	0.6383
TRIPLEX	0.7907	0.7394	0.5766	0.9311	1.3456	0.6428	0.7919	0.7432	0.5709	0.8768	1.2887	0.6533
Stem	**0.8298**	**0.7726**	**0.5984**	**0.7547**	**1.0742**	**0.0693**	**0.8365**	**0.7651**	**0.5748**	**0.7447**	**0.9631**	**0.0862**

Recover variations in gene expression:

Marker gene visualization:

Ablation Study

Different datasets from different studies, species, organs, health conditions:

❖ Kidney Visium dataset: 23 kidney sections from 22 individuals, containing three different health conditions
❖ Human prostate cancer (PRAD) Visium dataset
❖ Healthy mouse brain Visium dataset

Stem maintains its top performance among existing algorithms regardless of the differences in biology.

Choice of sampling statistics:
- Overall, the sample mean achieves the best numerical performance, and thus we choose it to be the predicted gene expression value.

Choice of pathology foundation models:
- Larger pathology foundation models do not necessarily imply a better performance for Stem.
- CONCH+UNI achieves the best performance in our experiments.

Power of histology image patch encoder:
Retrieving the nearest neighbors from the training dataset based on the image features output from those histology image patch encoders, and simply taking the average of the gene expressions from these selected neighbors as the final prediction for the test patch.

Model	HMHVG					
	PCC-10↑	PCC-50↑	PCC-200↑	MAE↓	MSE↓	RVD↓
ResNet18	**0.4795**	**0.3999**	0.2297	1.0124	1.7687	0.4064
CONCH	0.3824	0.3250	0.2442	0.9805	1.5618	**0.2687**
UNI	0.4328	0.3779	**0.2909**	**0.9012**	**1.3785**	0.3599
CONCH + UNI	0.3954	0.3417	0.2547	0.9301	1.4506	0.3269
Stem (CONCH + UNI)	0.5893	0.5332	0.4257	0.8792	1.3513	0.0751

This comparison indicates that Stem achieves a nontrivial improvement on the top of these image encoders.

Reference

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3):850–862, 2024.
Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. Nature Medicine, 30(3):863–874, 2024.
William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023.