

Integrating Histology Imaging and Spatial Transcriptomics through Low-Dimensional Embeddings to Predict Gene Expression Patterns



Gaurav Anand ¹ Lauren Cooper ² Susie Avagyan ³

¹Stanford Center for Professional Development ²Stanford School of Mechanical Engineering ³Stanford School of Medicine

Introduction

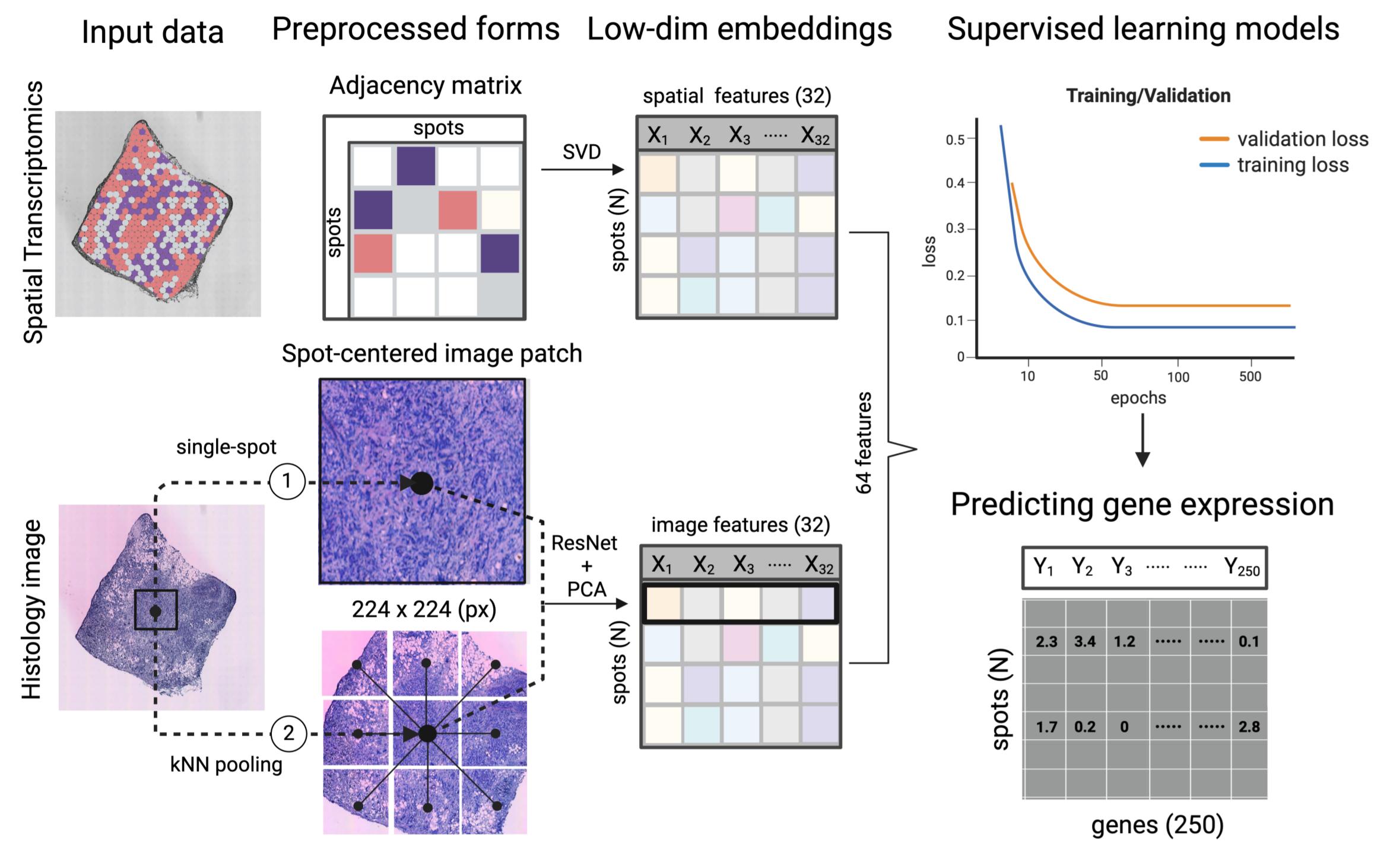


Figure 1. Overview of the framework: From input ST and histology imaging data to prediction of 250 target genes.

(a) Dotted lines represent alternative preprocessing routes for spot-centered image patches.

The study aims to leverage histology images and spatial adjacency information to predict spot-level gene expression patterns from readily available histology images as a cost-effective alternative to spatial transcriptomics (ST). Utilizing a combination of ResNet50 for image feature extraction and various regression and deep learning models (including linear regression, negative binomial regression, simple neural networks, and graph neural networks), we constructed a predictive framework which infers gene expression, leveraging low-dimensional embeddings of histology images and spatial information. The results demonstrated that for a log-normalized target, linear regression models with regularization and simple neural network models showed promising performance, suggesting the advantage of lower complexity models when using embedded features. However, moderate performance on the test set indicates the need for further model refinement and exploration of advanced techniques to fully capture the complex relationships between histology images and gene expression patterns.

Dataset and Features

Our data was obtained from the "Human breast cancer in situ capturing transcriptomics" dataset [2].

- **Dataset**
 - 23 breast cancer patients
 - 68 breast tissue sections
 - ST data of 30,612 spots 1 spot $\approx 100 \mu\text{m} \approx 224 \times 224 \text{ px}$
- **Features**
 - total 64 embedded features (Fig. 1)
 - 32 image features on each 224x224 image patch (Fig. 1a)
 - 32 spatial features on each spot
- **Target**
 - preselected 250 genes with the highest variance and mean expression levels (Fig.2)

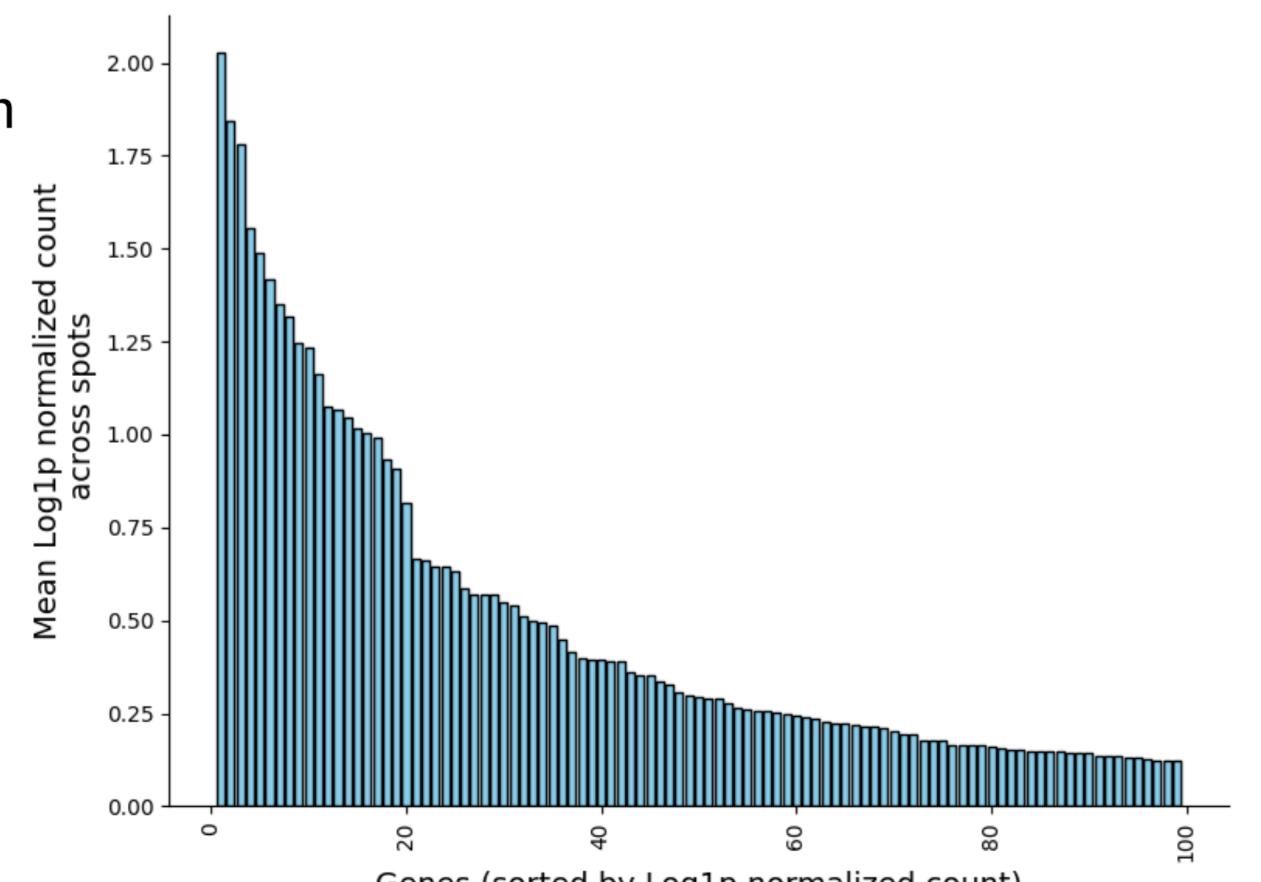


Figure 2. Log1p normalized mRNA Counts for top 100 genes

Models

Investigated several modeling approaches based on kNN-pooling and single-spot image preprocessing (Fig 1a)

- Negative Binomial
 - Fitted on raw gene expression target
- Linear regression (on log-normalized target)
 - Simple Linear
 - Ridge
 - Lasso
- Simple 3-layer Neural Network
 - Two dropout layers for regularization
 - Uses low-dimensional encoding to reduce feature space
- Graph Neural Network
 - Utilizes full feature space without dimensionality reduction
 - Two graph convolutional layers utilizing batch normalization and dropout for regularization

Results

Et rutrum ex euismod vel. Pellentesque ultricies, velit in fermentum vestibulum, lectus nisi pretium nibh, sit amet aliquam lectus augue vel velit. Suspendisse rhoncus massa porttitor augue feugiat molestie. Sed molestie ut orci nec malesuada. Sed ultricies feugiat est fringilla posuere.

Table 1: Modeling Metric Comparison across final model suite (truncated for space)

Name	Split	PCC	MSE
Simple Neural Network	Train	0.3023	0.1884
	Valid	0.0391	0.3229
	Test	0.0639	0.3028
Linear Regression Lasso	Train	0.1607	0.2344
	Valid	0.1759	0.2751
	Test	0.1243	0.2881
KNN Pooled GNN	Train	0.14297	0.2363
	Valid	-0.0260	0.3036
	Test	0.02887	0.2960
KNN Linear Regression Ridge	Train	0.32554	0.1766
	Valid	0.25974	0.2243
	Test	0.06064	0.3600
NB Alpha Apriori Estimation	Train	0.2130	128.6232
	Valid	0.0568	8.0991
	Test	0.0420	26.1523
KNN NB Alpha Apriori Estimation	Train	0.2838	588.8546
	Valid	0.0687	6.1521
	Test	0.0273	49.3884

Figure 3. Another figure caption.

Results

Linear Regression

We found that regularized linear regression performed the best, as our Lasso regularization yielded the best results.

NB Regression

Results, cont.

Simple Neural Network

Graph Neural Network

Discussion

Our models

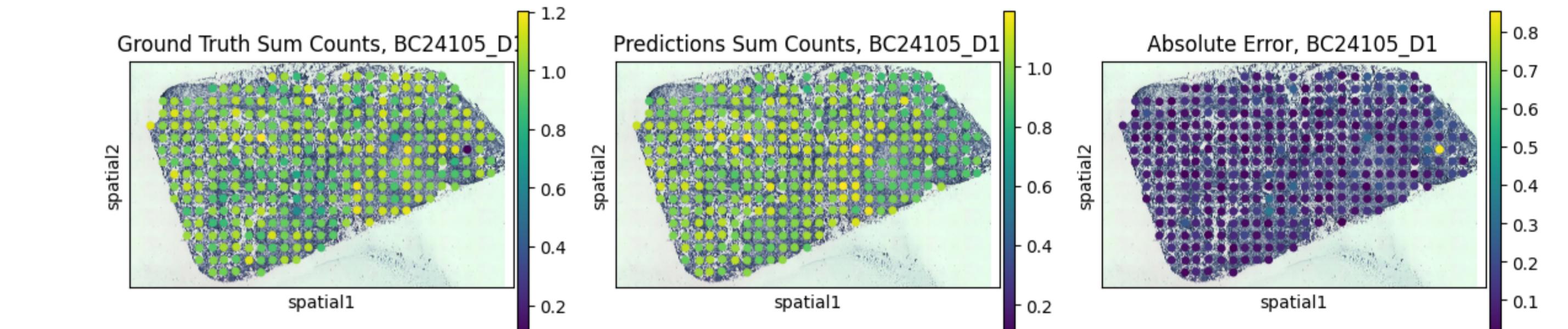


Figure 4. Caption

Model Performance

Image of

Another heading inside a block

Sed augue erat, scelerisque a purus ultricies, placerat porttitor neque. Donec $P(y | x)$ fermentum consectetur $\nabla_x P(y | x)$ sapien sagittis egestas. Duis eget leo euismod nunc viverra imperdiet nec id justo.

Future Work

Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Phasellus libero enim, gravida sed erat sit amet, scelerisque congue diam. Fusce dapibus dui ut augue pulvinar iaculis.

First column	Second column	Third column	Fourth
Foo	13.37	384,394	α
Bar	2.17	1,392	β
Baz	3.14	83,742	δ
Qux	7.59	974	γ

Table 1. A table caption.

Donec quis posuere ligula. Nunc feugiat elit a mi malesuada consequat. Sed imperdiet augue ac nibh aliquet tristique. Aenean eu tortor vulputate, eleifend lorem in, dictum urna. Proin auctor ante in augue tincidunt tempor. Proin pellentesque vulputate odio, ac gravida nulla posuere efficitur. Aenean at velit vel dolor blandit molestie. Mauris laoreet commodo quam, non luctus nibh ullamcorper in. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos.

Nulla varius finibus volutpat. Mauris molestie lorem tincidunt, iaculis libero at, gravida ante. Phasellus at felis eu neque suscipit suscipit. Integer ullamcorper, dui nec pretium ornare, urna dolor consequat libero, in feugiat elit lorem euismod lacus. Pellentesque sit amet dolor mollis, auctor urna non, tempus sem.

References

[1] Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379–423, 1948.