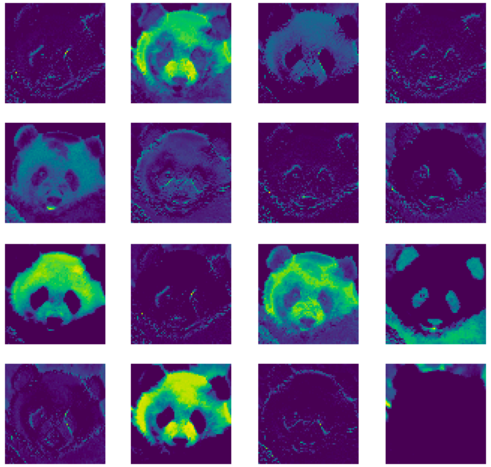
Interpretable Neuron Structuring with Graph Spectral Regularization

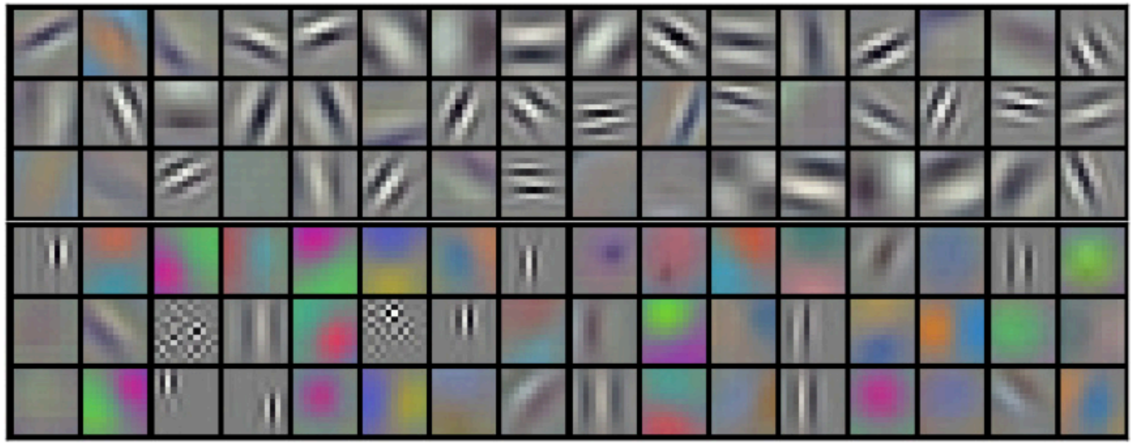
Alexander Tong, David van Dijk, Jay S. Stanley, Matthew Amodio, Kristina Yim, Rebecca Muhle, James Noonan, Guy Wolf, and Smita Krishnaswamy

Yale

Convolutional NN filter interpretability

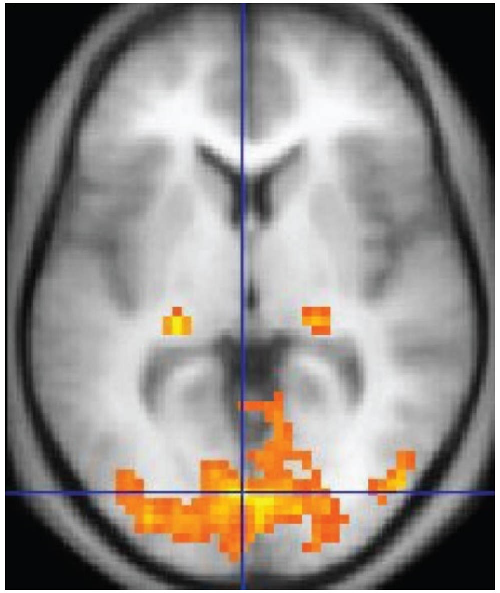
- Filter maps
- Activation maps
- Gradient based methods [Olah et al. 2017]
- Up-convolutional net [Dosovitskiy and Brox 2016]

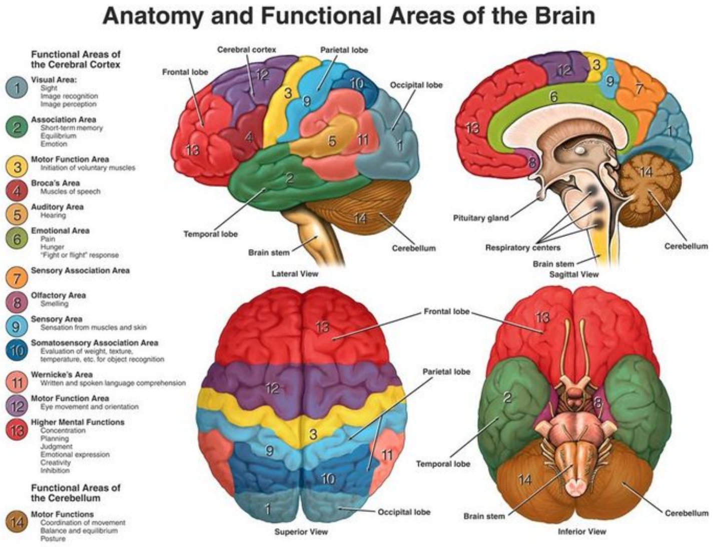




Krishevsky et al. 2012, Zhang et al. 2018 CVPR

Can we make interpretable activation maps for fully-connected NNs?





Analogy to real neural networks

- Often preprocessed into “functional regions”
- X condition has activation / suppression in Y region
- We can gain a high-level understanding of real brains by summarizing 10^11 neurons into localized groups

Organizing layers with graph structure

Enforcing graph structure

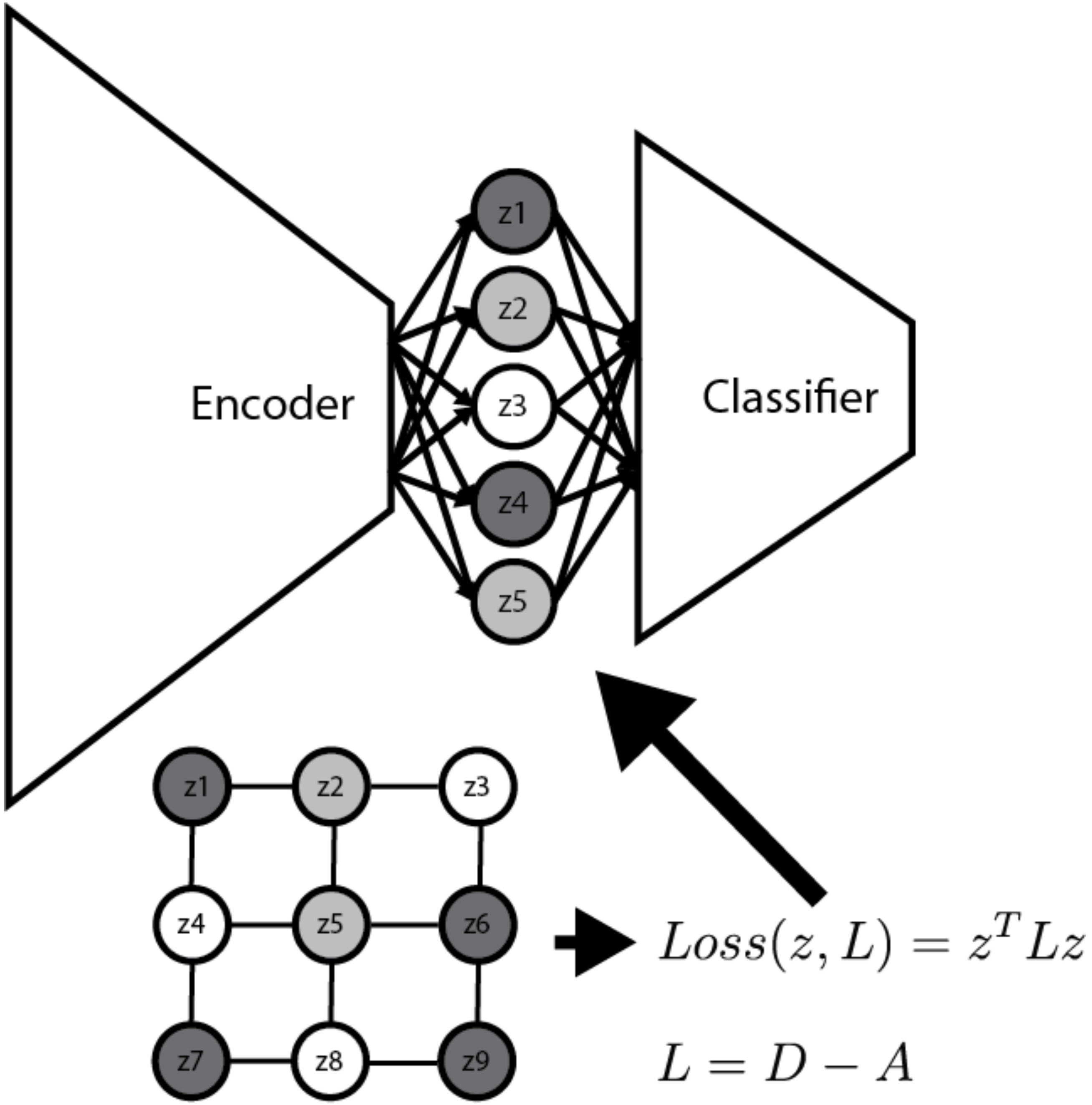
- Take a predefined graph and force activations to be smooth on that graph

Learning graph structure

- Simultaneously optimize the graph structure and activation smoothness

Enforcing a Grid Structure on MNIST

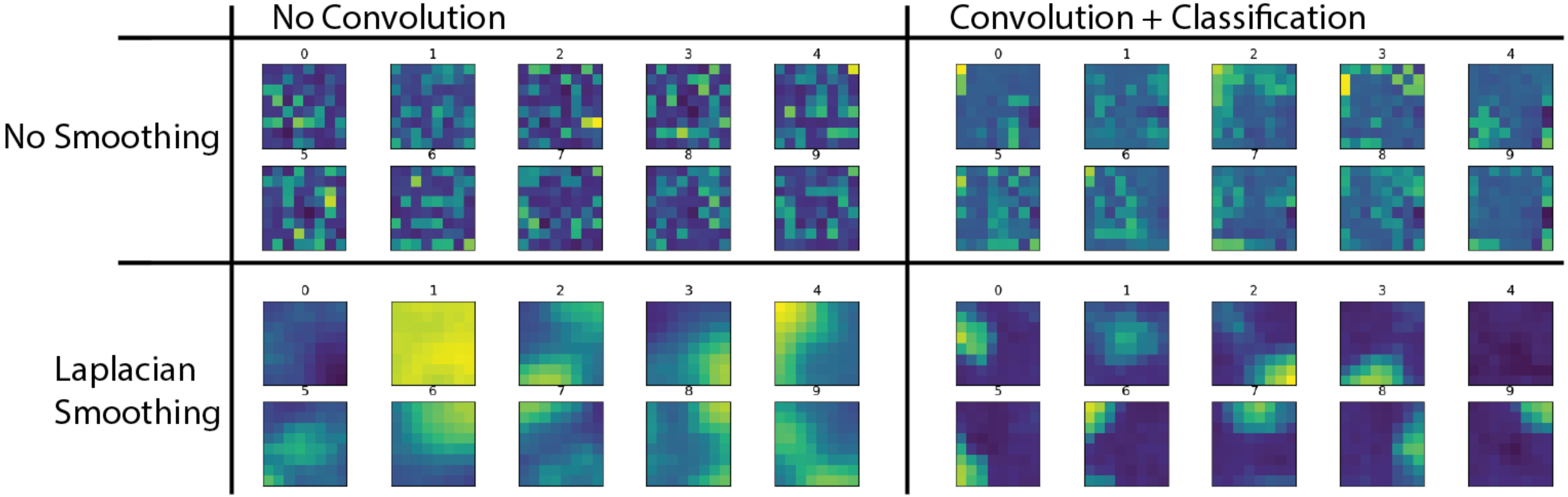
- MNIST classification with dense encoder
- 64 width layer enforcing an 8x8 grid structure
- Two methods
  - Convolutional classifier
  - Graph smoothing



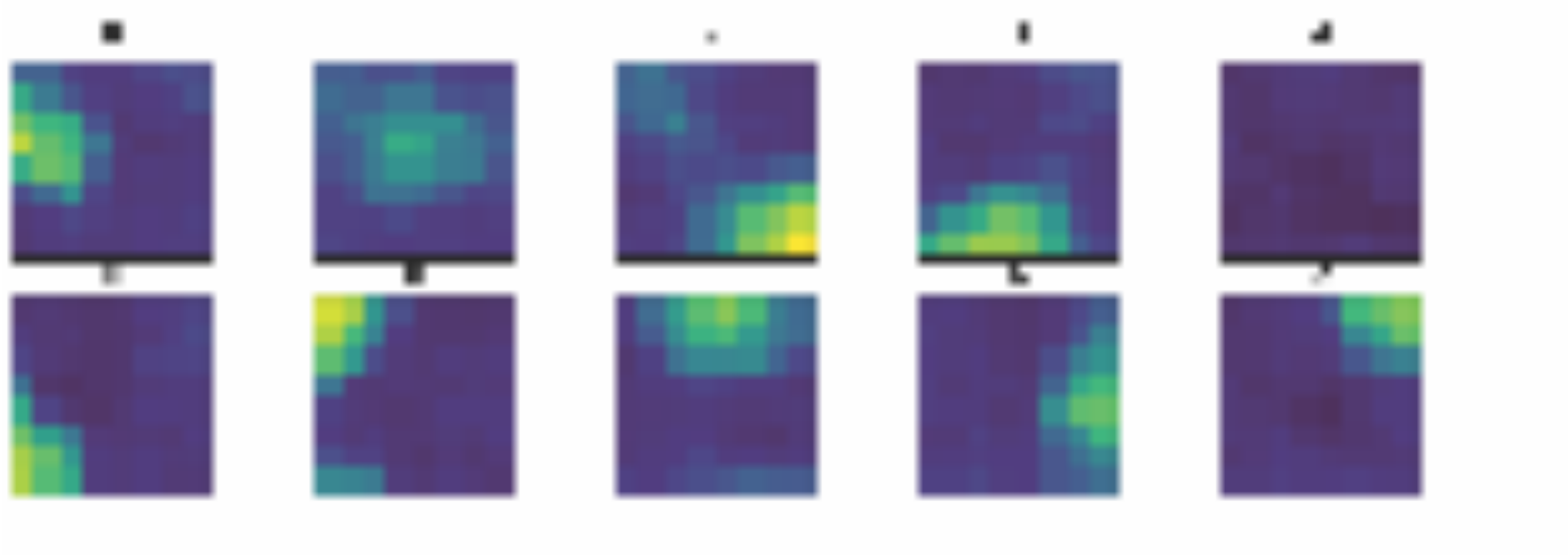
$$Loss(z, L) = z^T L z$$

$$L = D - A$$

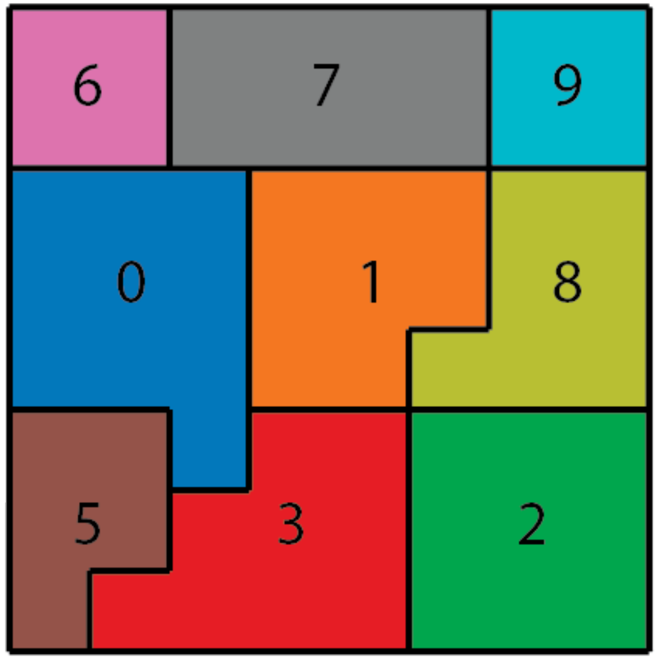
Activation Maps for MNIST

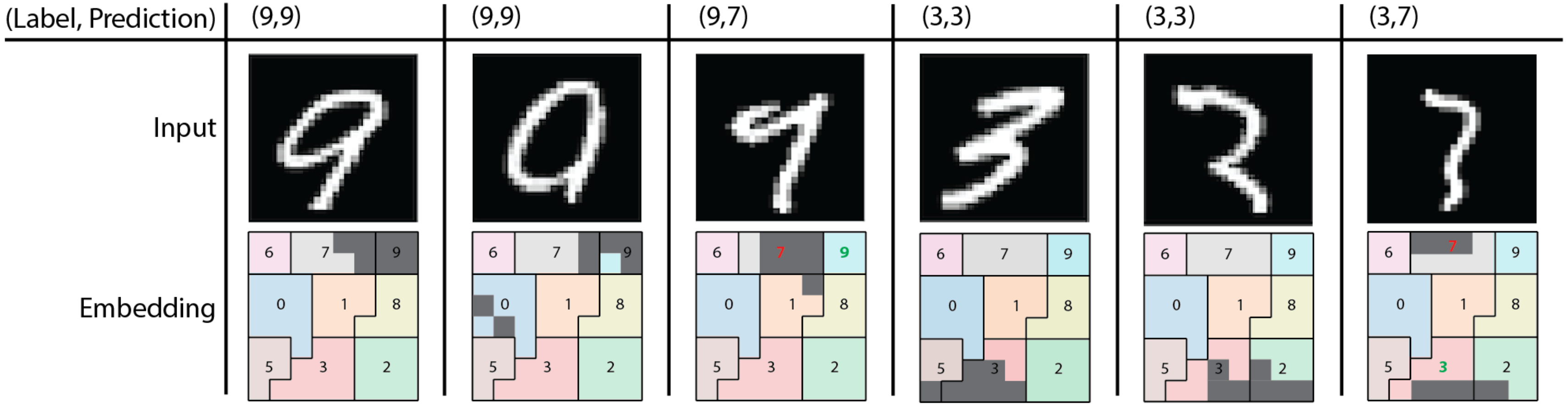


Convolution + Graph regularization



Segmentation

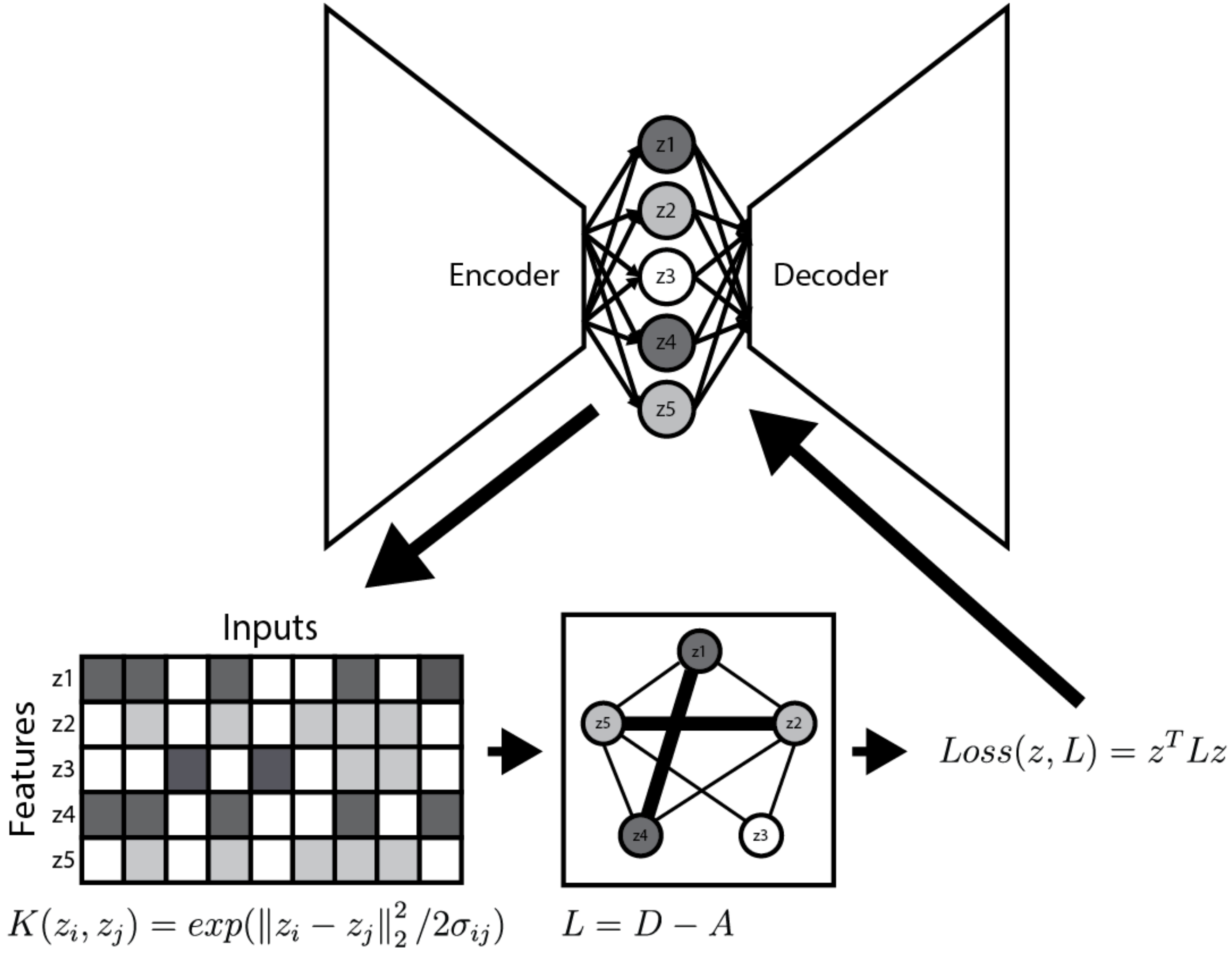




Learning a Graph Structure

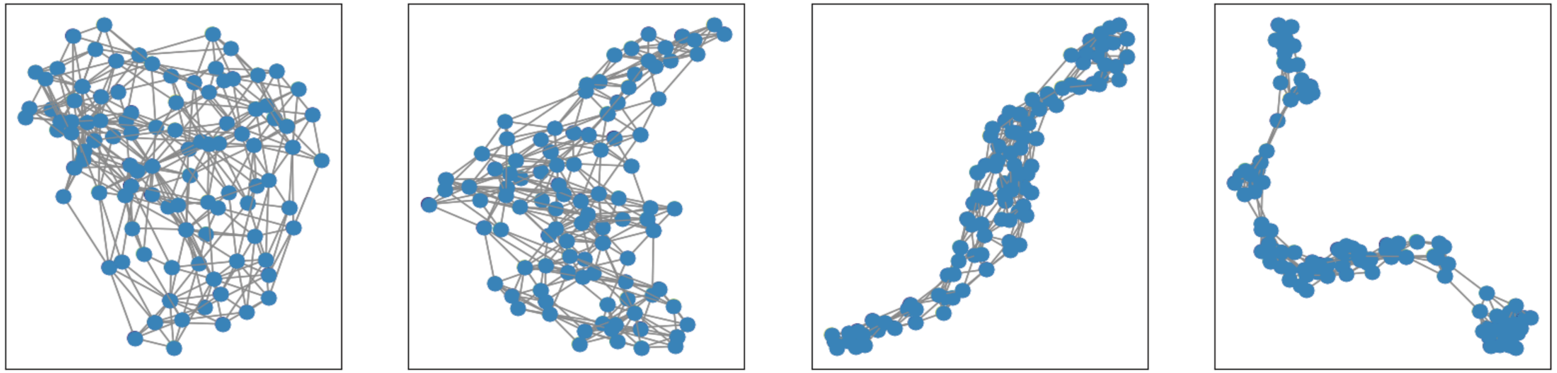
Repeatedly do:

- Create graph from gaussian kernel on activations
- Train for M steps with GSR loss

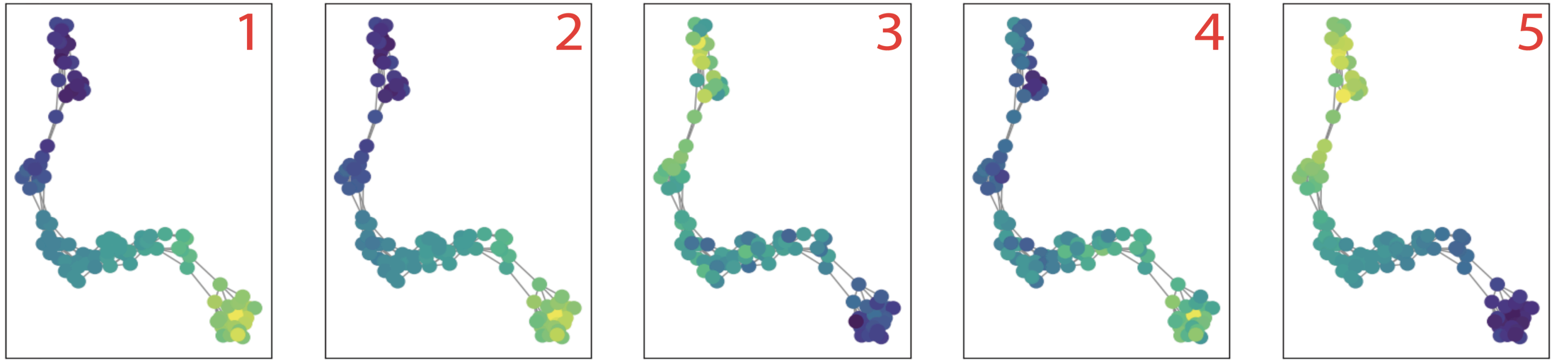


Learning the graph in a single-cell (cell X gene) dataset

a) Training Time →

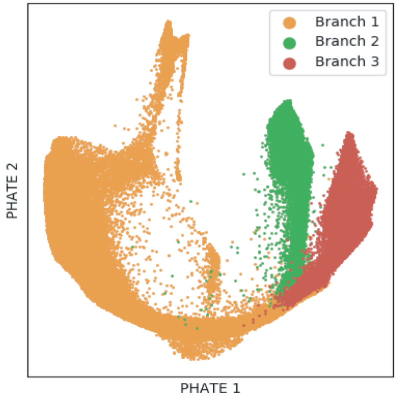


b)

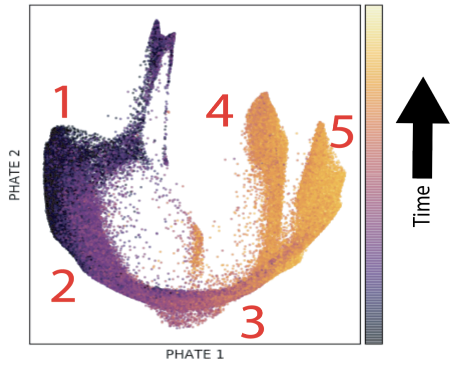


Extracted Graph Structure of Genes

c) Developing T-cells



d)



Visualization of cells

Setty et al. 2016

Summary

Fully connected layers have no natural coherent structure

Imposing a graph structure can create locality like a brain

Graph structure can be learned from the data

Acknowledgements

Colleagues

- Krishnaswamy Lab
- Noonan Lab

Funding

- IVADO
- Chan-Zuckerberg Initiative
- NIH

Lab Website: www.krishnaswamylab.org

Code: https://github.com/KrishnaswamyLab/GraphSpectralRegularization