UNIVERSITY OF TORONTO

Tutorial: Latent Space Interpretation

Foundation Models for Science Workshop

4 November 2025

Schmidt Sciences

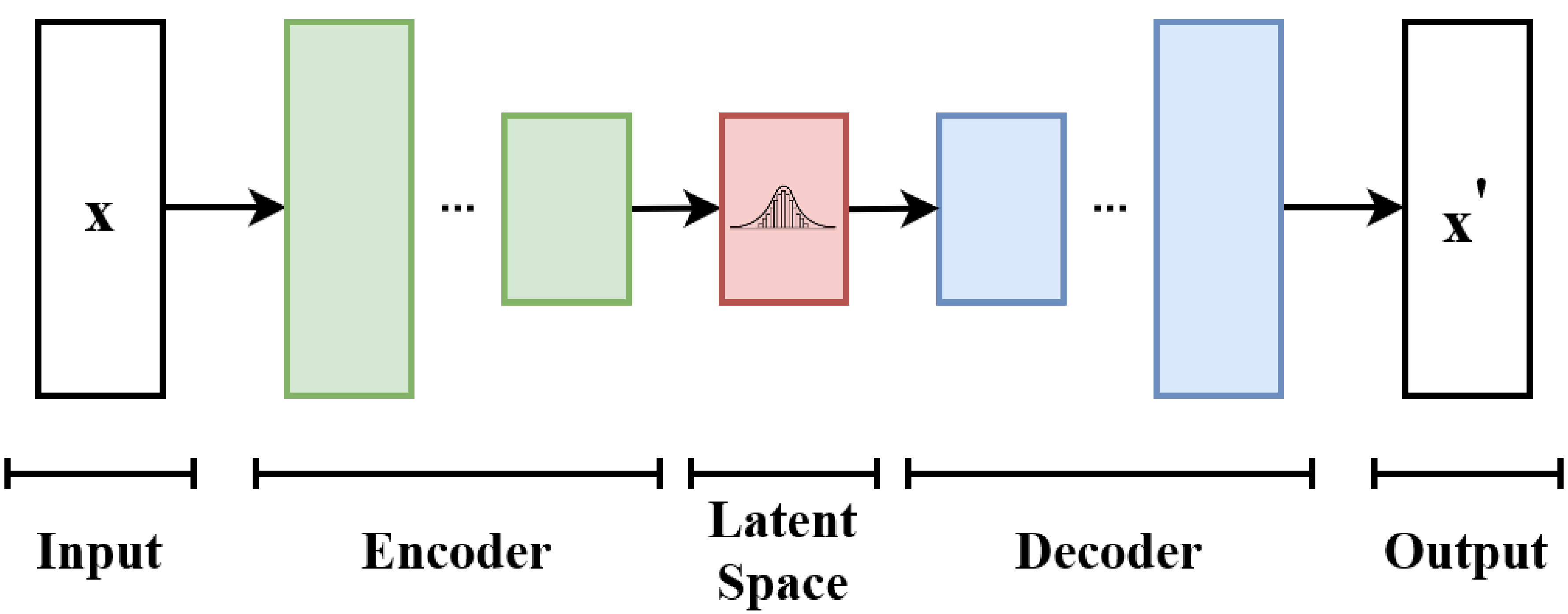
FOUNDATION MODELS for SCIENCE

Warning!

- Think about these slides as a place to start collecting search terms for later

What is a latent space?

- “latent” == hidden



https://en.wikipedia.org/wiki/Variational_autoencoder

Learning Outcomes

1. Extract and manipulate embeddings
2. Reduce high-dimensional embeddings to 2D for visualization
3. Quantify clustering quality using mutual information metrics
4. Optimize dimensionality reduction hyperparameters automatically
5. Analyze how features change across different layers of a transformer model
6. Interpret latent space structure

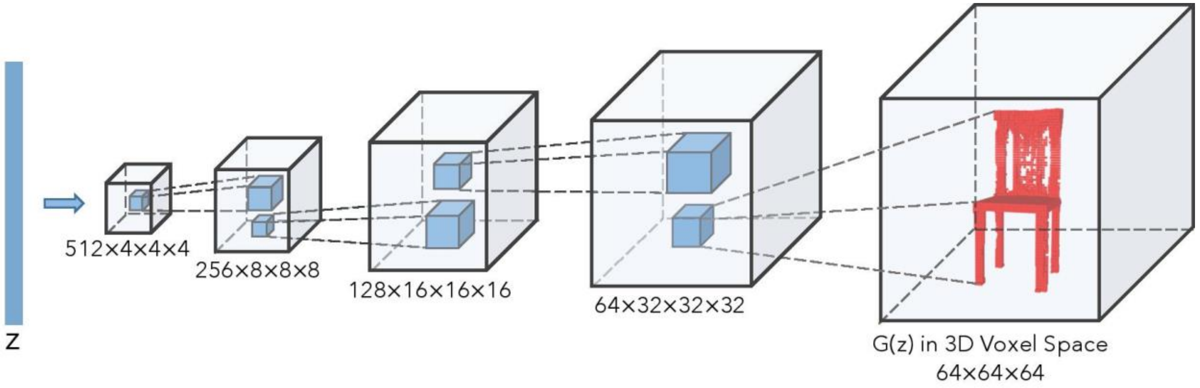
What questions can we answer with these methods?

- What information has the model learned?
- How is the model correlating information internally?

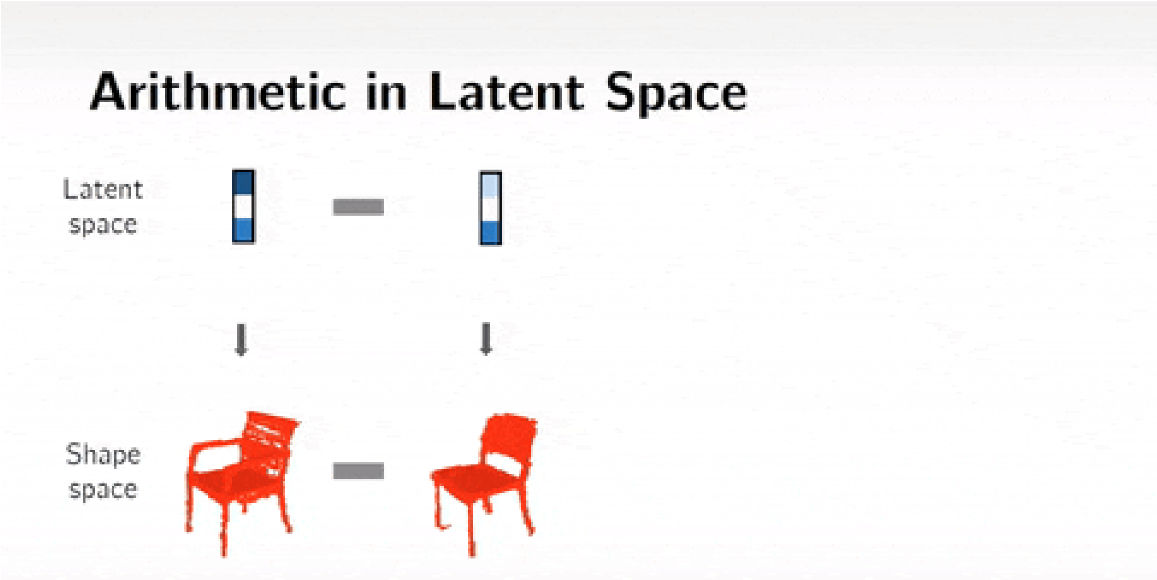
Example: A Model That Generates Chairs

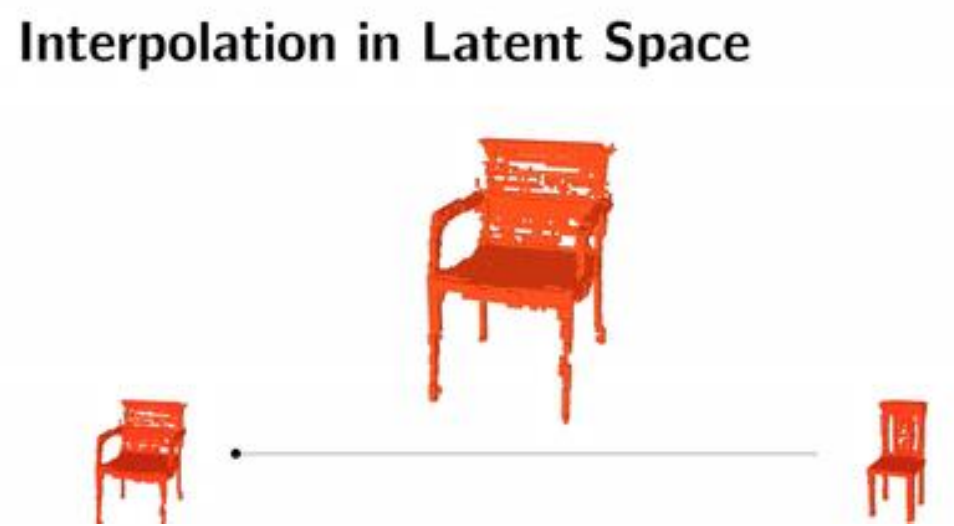
Wu, J., Zhang, C., Xue, T., Freeman, B., & Tenenbaum, J. (2016). **Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling**. *Advances in neural information processing systems*, *29*.

Latent vector, Z

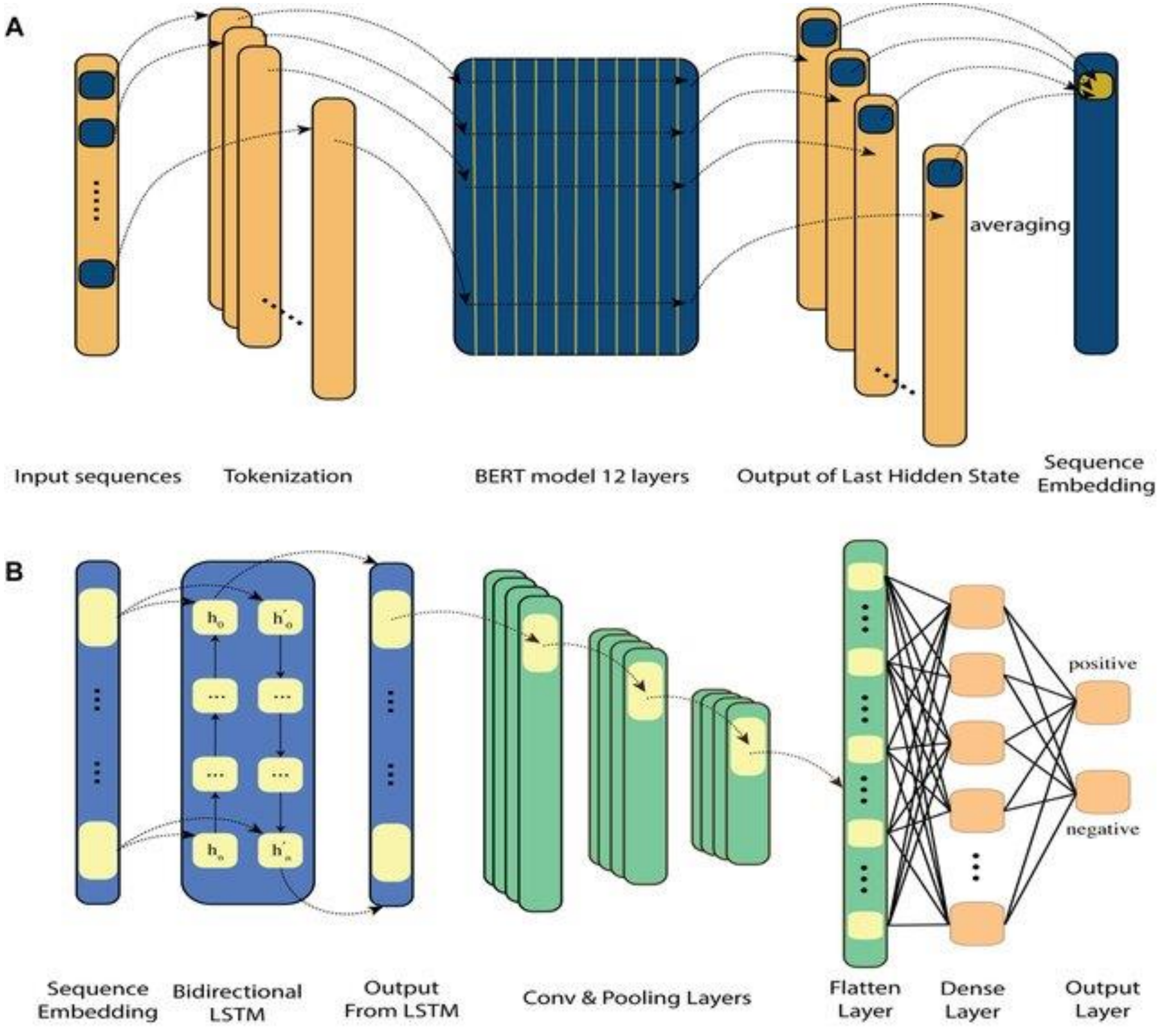


Output





1. Extract Embeddings



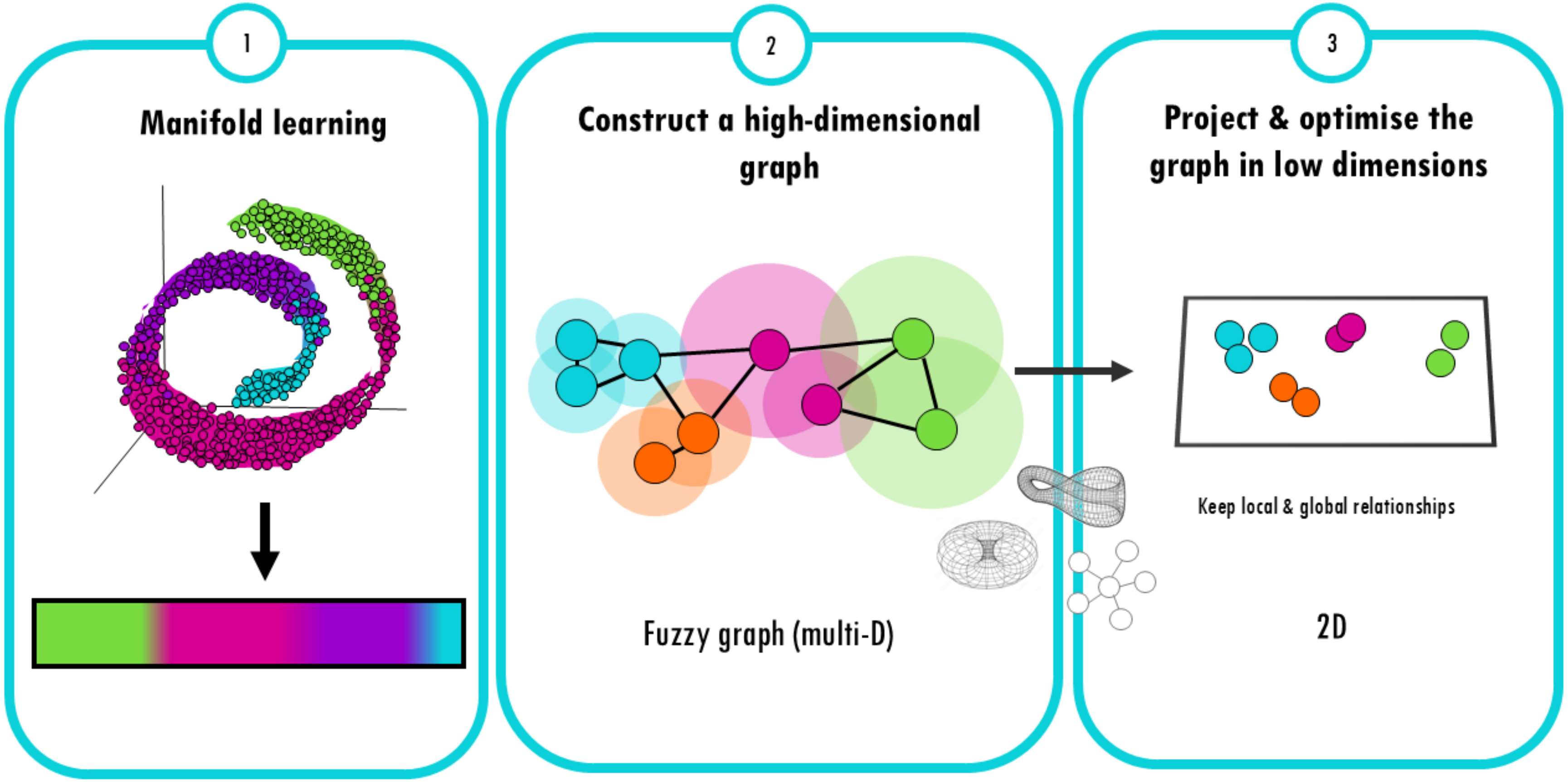
Which layer in the model to extract embeddings?

Things to consider:

- Do some layers have a particular function that you would like to probe? Use domain expertise
    - *Common practice: last layer before prediction*
- Some feature vectors are very large; are they computationally tractable for the implemented method?

Jiang, Hanli, et al. "SenSeqNet: A Deep Learning Framework for Cellular Senescence Detection from Protein Sequences." *bioRxiv* (2024): 2024-10.

2. Dimensionality Reduction – UMAP Algorithm



https://biostatsquid.com/umap-simply-explained/

2. Dimensionalty Reduction Cont.

- Shout out to other methods!
  - PCA
  - SVD
  - T-SNE
  - VAEs
  - Spectral Embedding (my favorite)
  - Hand shadows on the wall

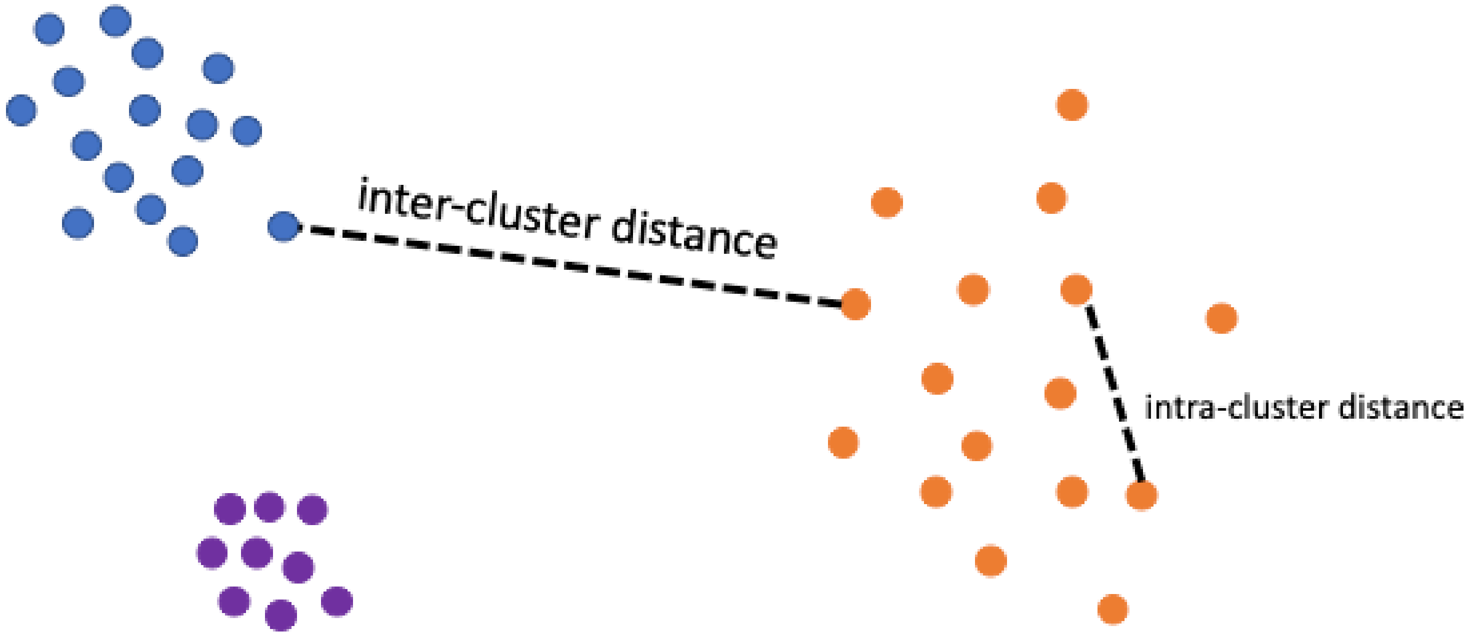
3. Clustering Methods

Clustering assigns labels to points

We can assign labels using different clustering methods

An ideal clustering method will *maximize information between the cluster label and the true label*

We can test different clustering methods to see which one works the best for our data



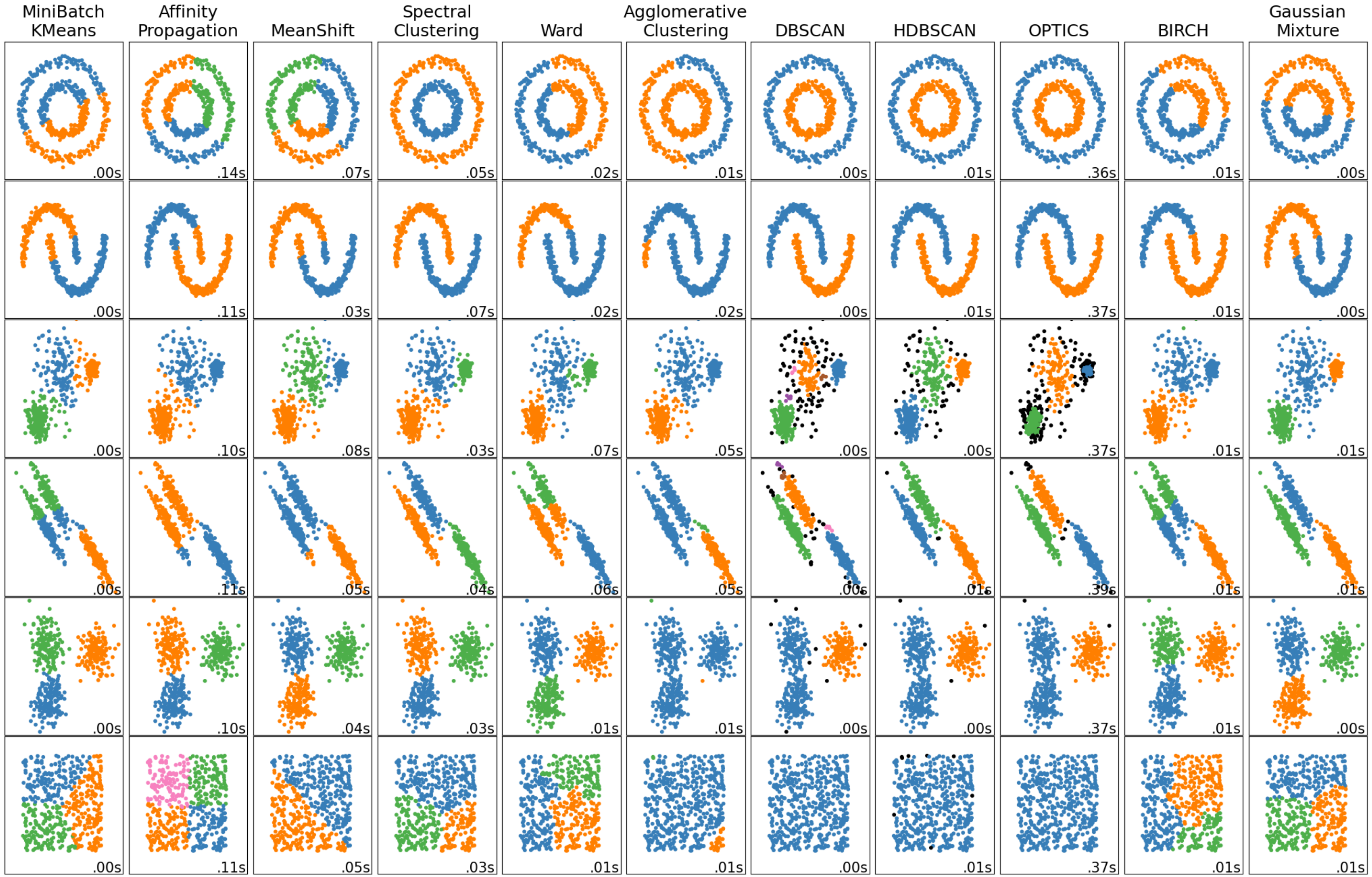
https://medium.com/data-science/three-performance-evaluation-metrics-of-clustering-when-ground-truth-labels-are-not-available-ee08cb3ff4fb

3. Clustering Methods Continued

Which of these clustering algorithms did the best?

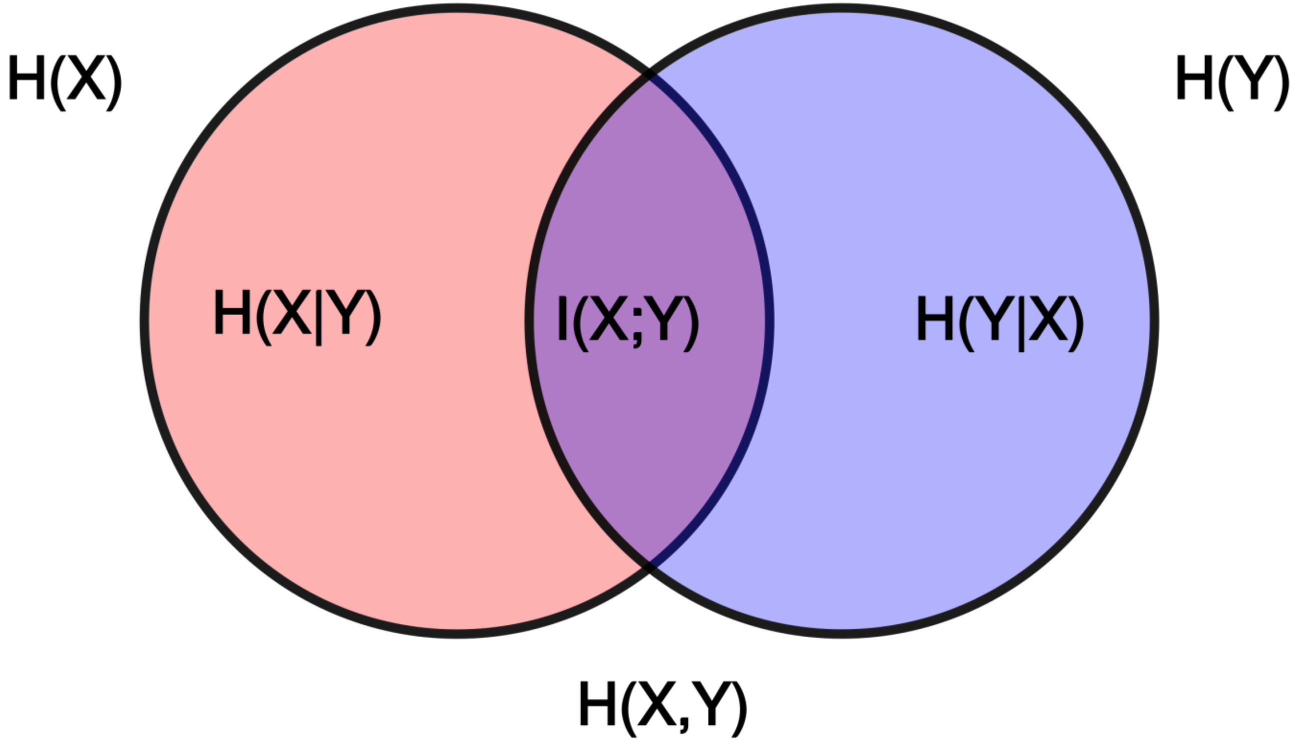
How do we know?

How can we quantify?



https://scikit-learn.org/stable/modules/clustering.html

3. Clustering Metric: Mutual Information



X: True labels
Y: Cluster labels

H(X): Entropy (information) of X
H(Y): Entropy (information) of Y

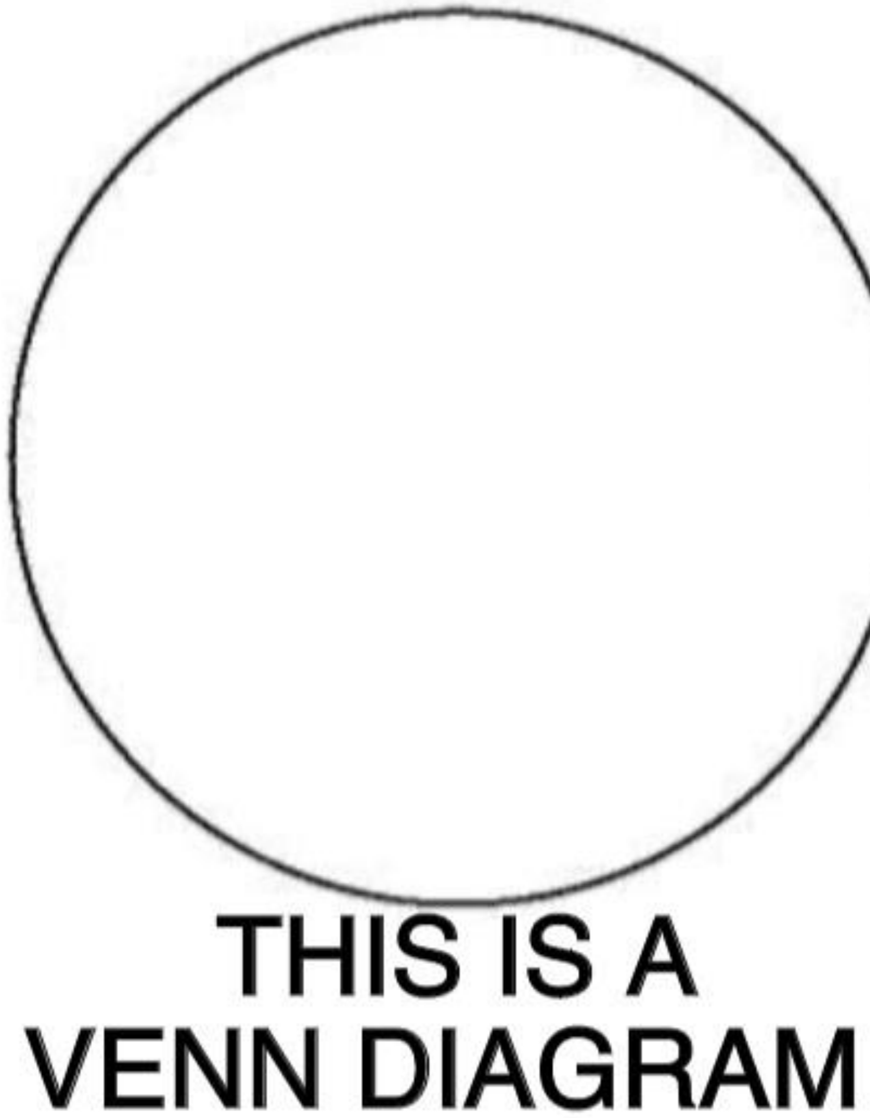
H(X|Y), H(Y|X): Conditional entropies

I(X; Y): Mutual Information

We want MI = 1

For other clustering metrics, google: ”scikit learn clustering-performance-evaluation”

https://en.wikipedia.org/wiki/Mutual_information


THIS IS A
VENN DIAGRAM

Learning Outcomes

~~1. Extract and manipulate embeddings~~

~~2. Reduce high-dimensional embeddings to 2D for visualization~~

~~3. Quantify clustering quality using mutual information metrics~~

4. Optimize dimensionality reduction hyperparameters automatically

5. Analyze how features change across different layers of a transformer model

6. Interpret latent space structure

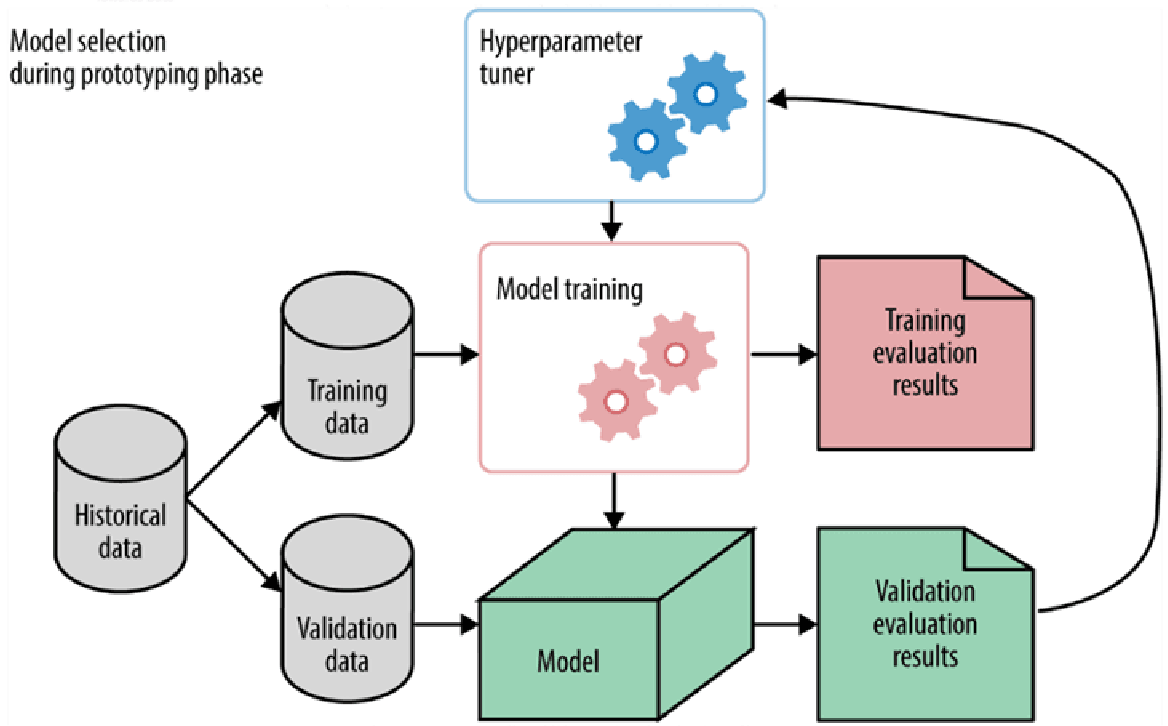
4. Optimize dimensionality reduction hyperparameters automatically

- Basic UMAP hyperparameters
  - Number of neighbors
  - Minimum distance between neighbors
  - Number of components after dimensionality reduction
  - How distance is measured (metric)
  - https://umap-learn.readthedocs.io/en/latest/parameters.html

- Use an optimization algorithm to find the best combination of hyperparmeters

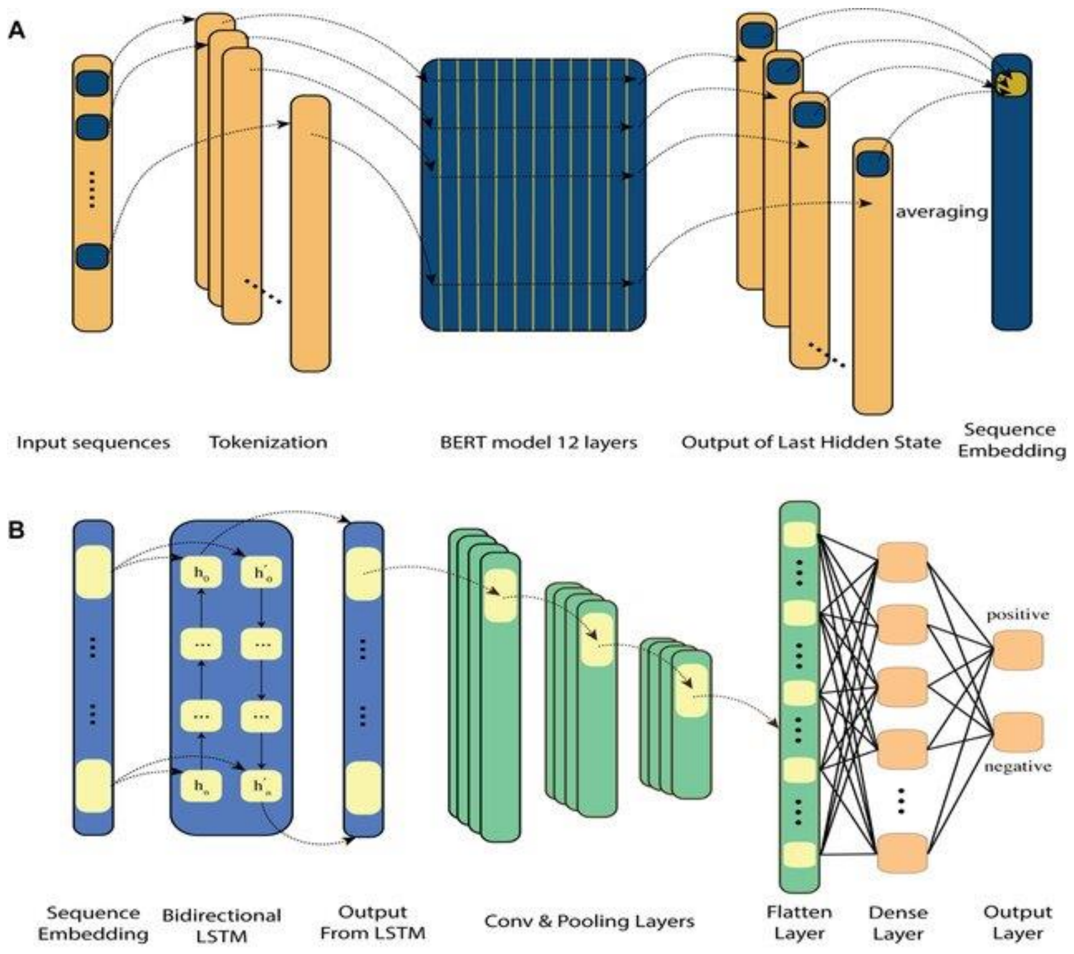
- “Best” means “maximizes MI”

- Sometimes computationally expensive, but definitely worth it



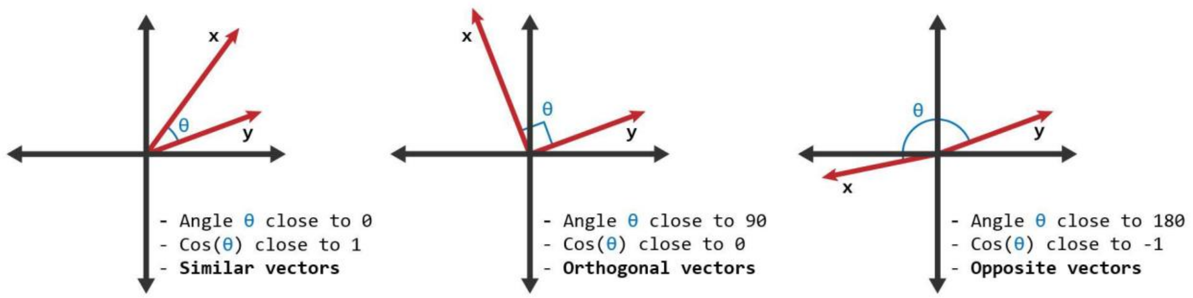
https://www.almabetter.com/bytes/articles/optuna-guide

5. Analyze how features change across different layers



This lets us re-evaluate an early decision about where to extract embeddings

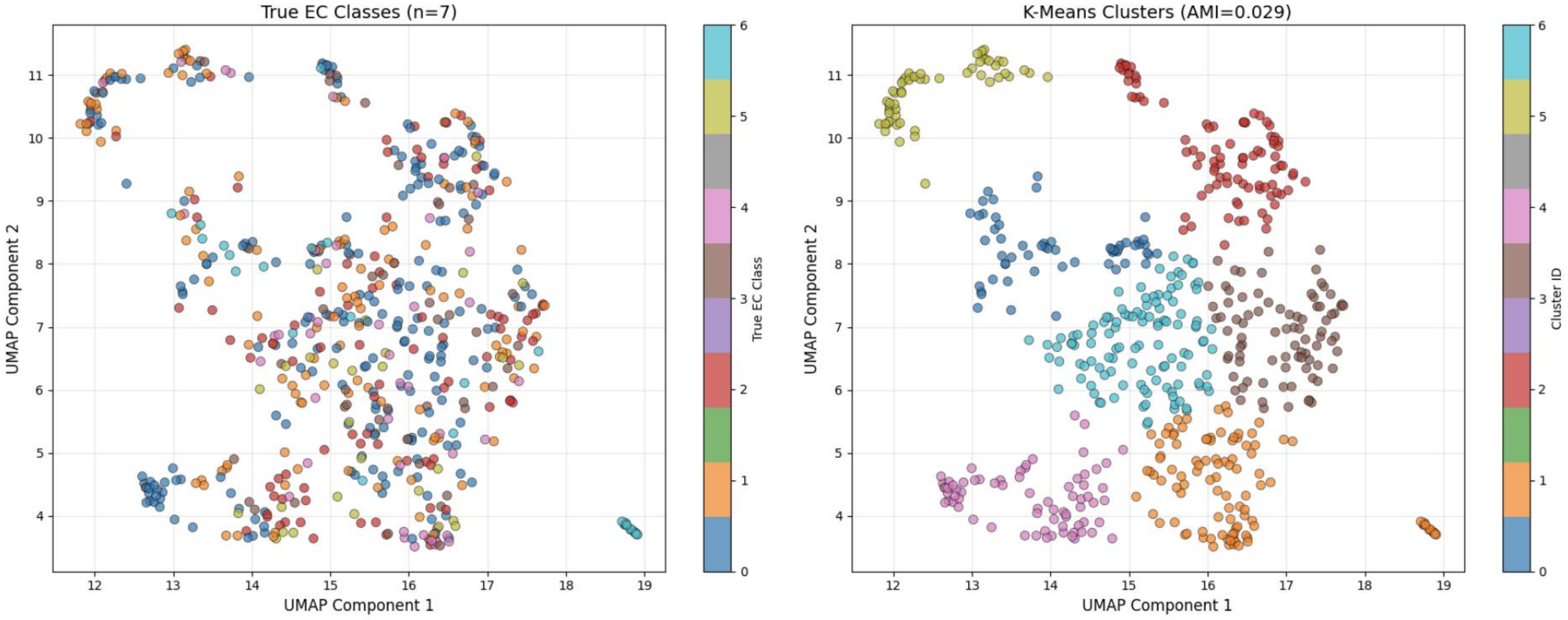
It also lets us see where representations stop changing in the model



https://www.learndatasci.com/glossary/cosine-similarity/

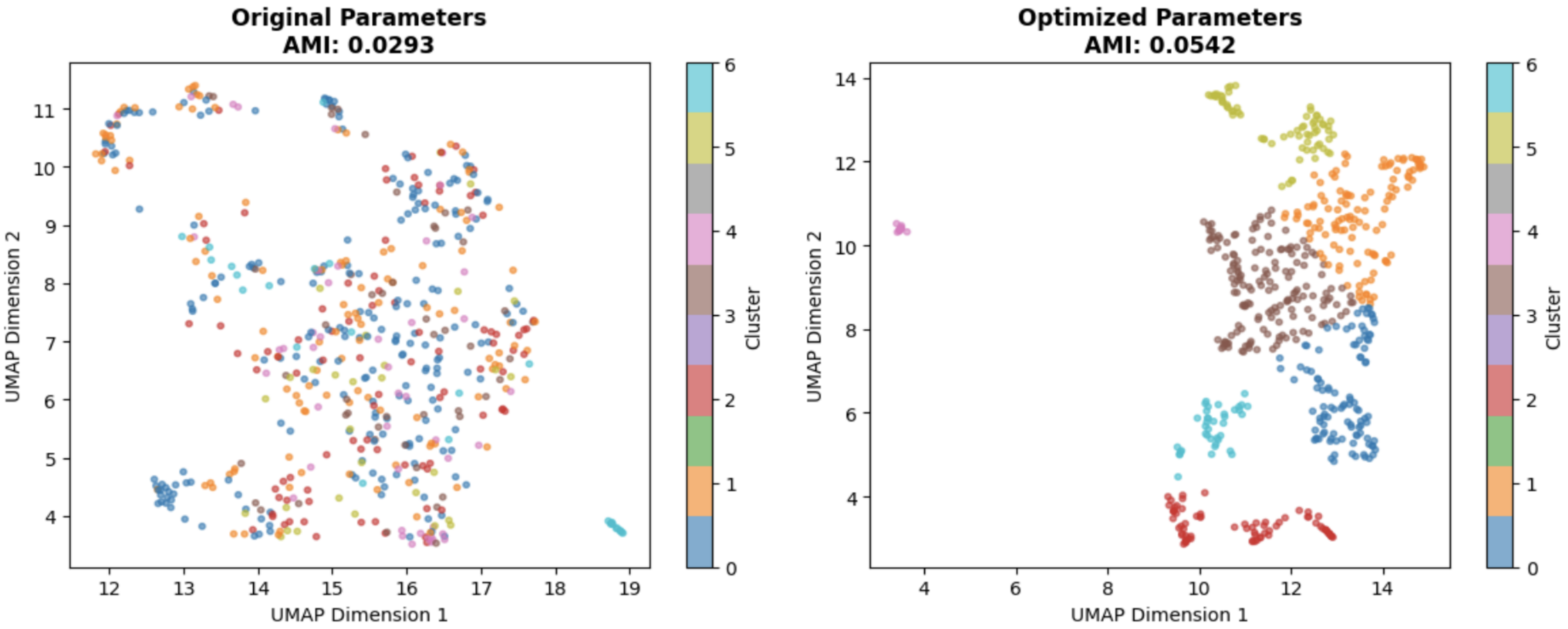
Jiang, Hanli, et al. "SenSeqNet: A Deep Learning Framework for Cellular Senescence Detection from Protein Sequences." *bioRxiv* (2024): 2024-10.

6. Interpreting latent space structure



<u>Question: Is this a good clustering result? Is the Mutual Information with the true labels high or low?</u>

6. Interpreting latent space structure

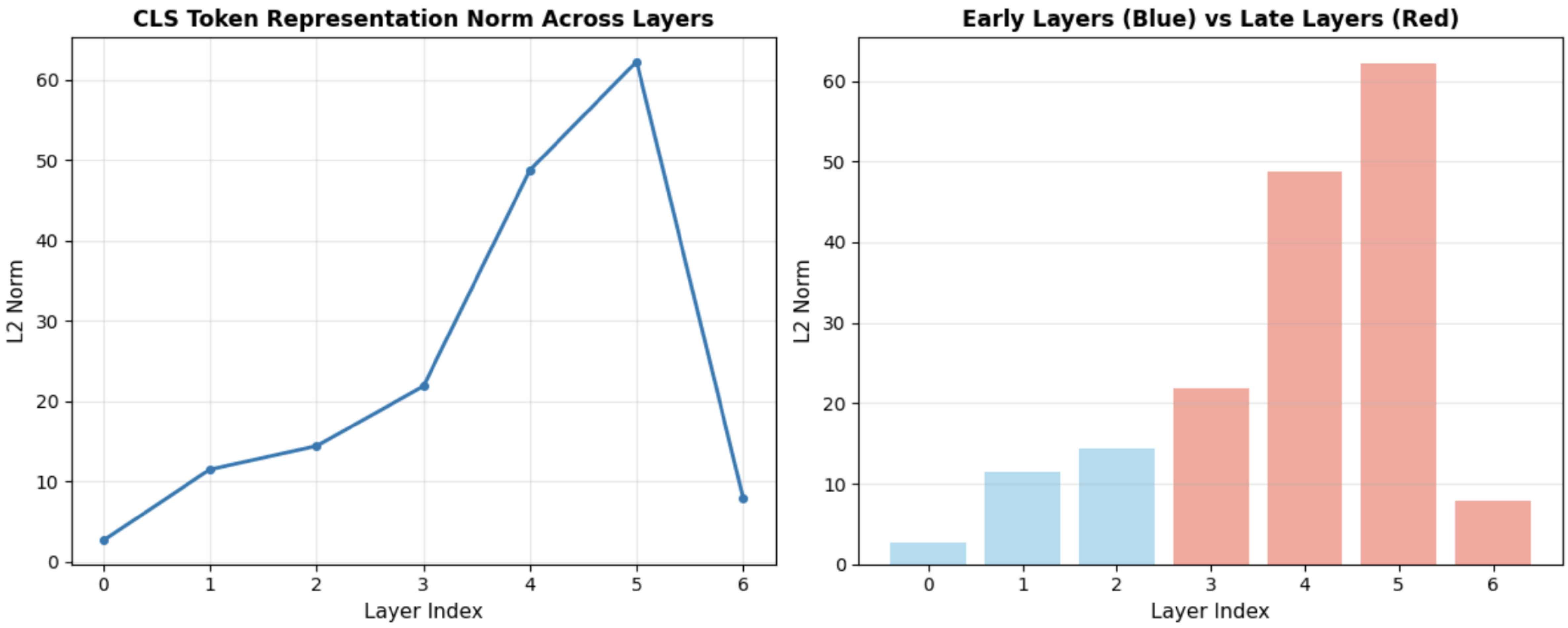


<u>Question: Is this a good clustering result? Is the Mutual Information with the true labels high or low?</u>

6. Interpreting latent space structure

- What does this tell us about our model?
- Is this measurement reliable? Why or why not?
- What can we do to change the results?

6. Interpreting latent space structure



<u>Question: Between which two layers is the representation changing the most?</u>

FOUNDATION MODELS
for SCIENCE