# Capturing biological patterns from gene expression data of PCOS using unsupervised dimensionality reduction algorithms

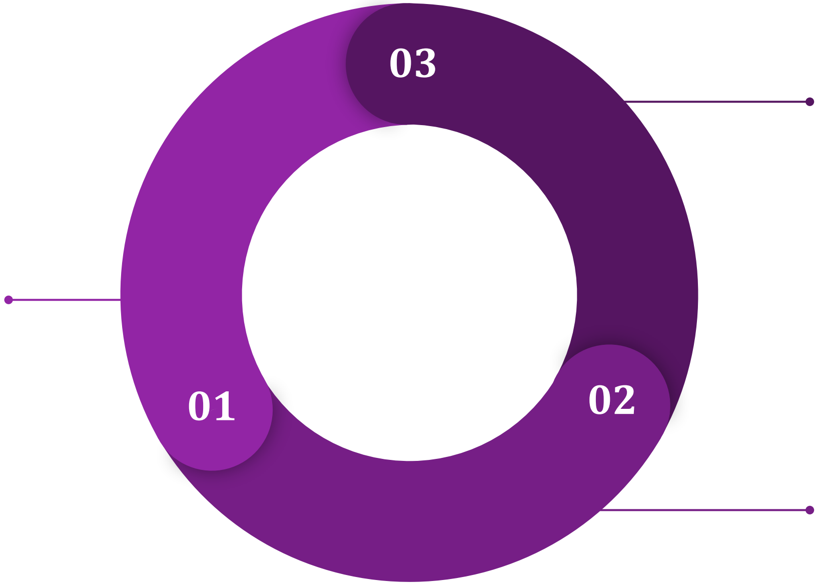
CS6024 | Algorithmic Approaches to Computational Biology | Course Project

N Sowmya Manojna (BE17B007)
A V Lakshmy (CS16B101)

December 14, 2020

# Introduction

**Gene expression analysis** is a powerful technique with many applications:



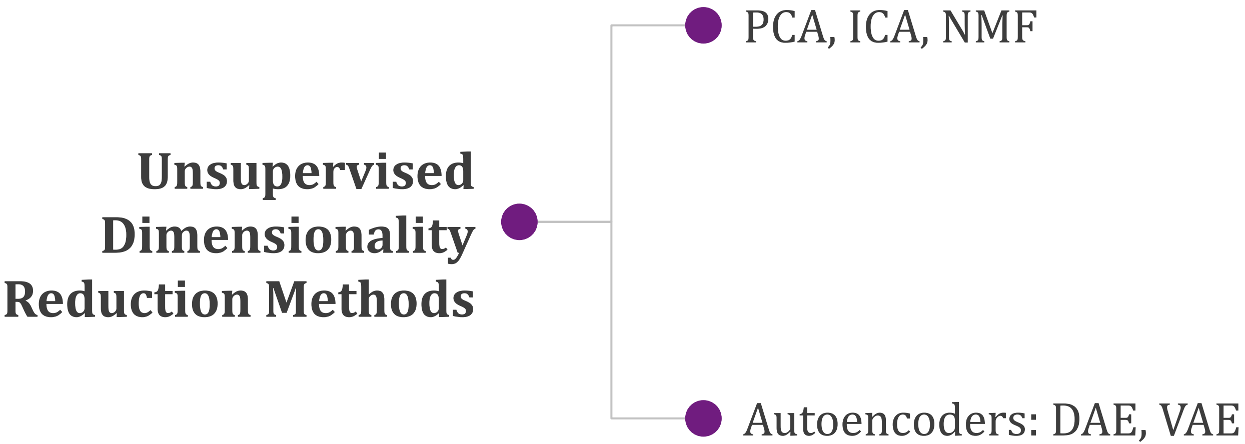
**Predicting onset of diseases**

**Evaluating efficiency of drugs**

**Studying effects of mutations**

# Background

**Gene expression data analysis is computationally expensive!!**



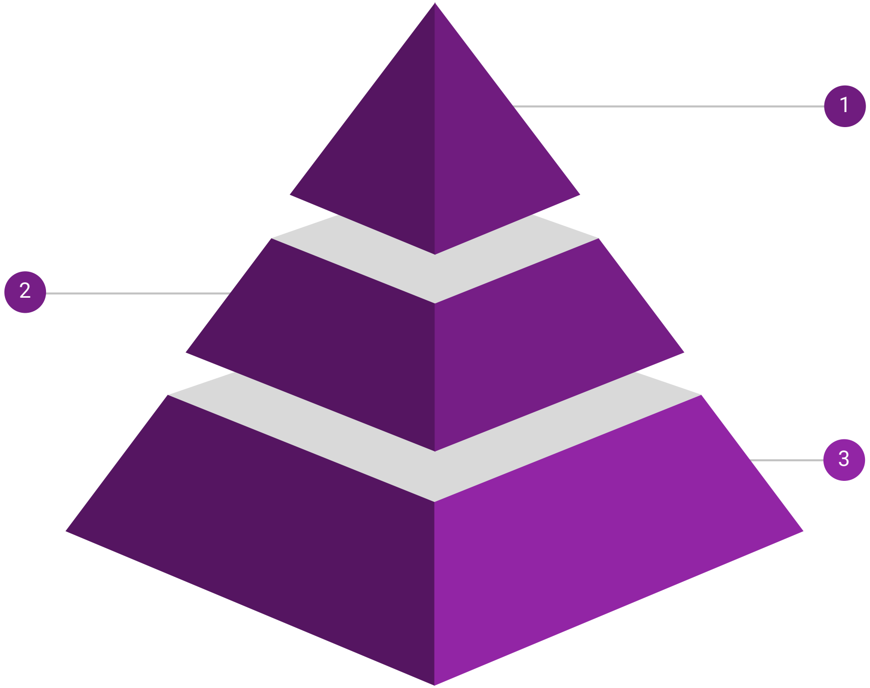
**Unsupervised Dimensionality Reduction Methods**

- PCA, ICA, NMF
- Autoencoders: DAE, VAE

# Problem Statement

*“Capturing biological patterns from gene expression data of Polycystic Ovarian Syndrome (PCOS) using unsupervised dimensionality reduction algorithms, particularly autoencoders”*

# What is PCOS? Why PCOS?

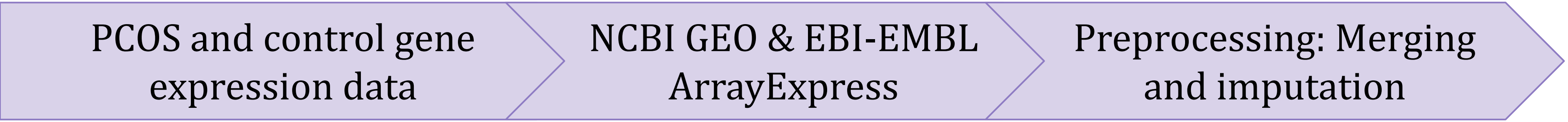


Endocrine disorder that affects 5-20% of women of reproductive age

No precise aetiology identified, but studies suggest genetic component

Prior studies use gene expression analysis with PCA

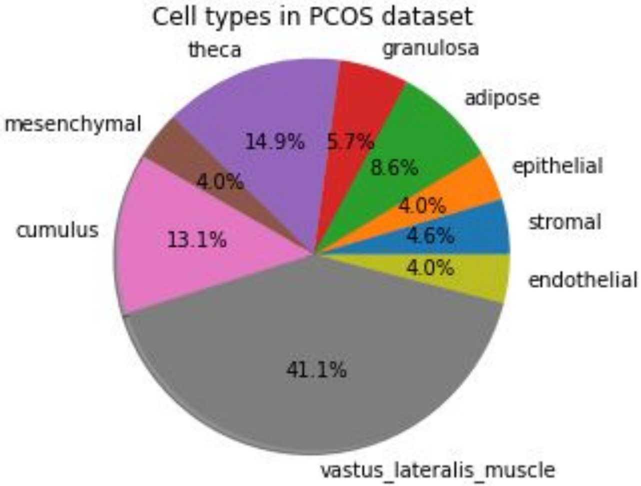
# Datasets



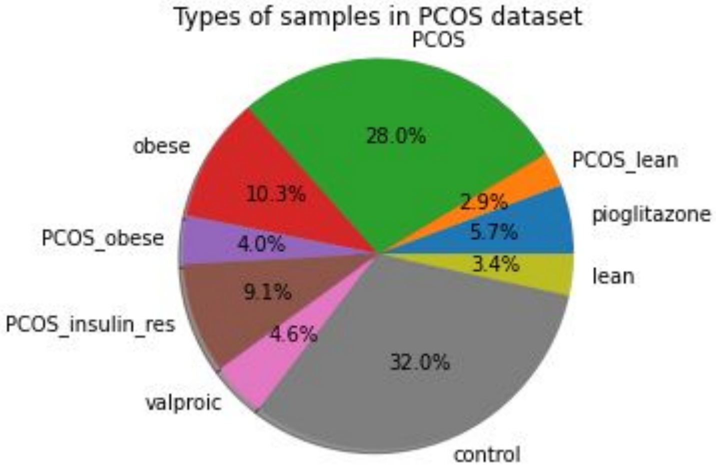
PCOS and control gene expression data

NCBI GEO & EBI-EMBL ArrayExpress

Preprocessing: Merging and imputation



Cell types in PCOS dataset



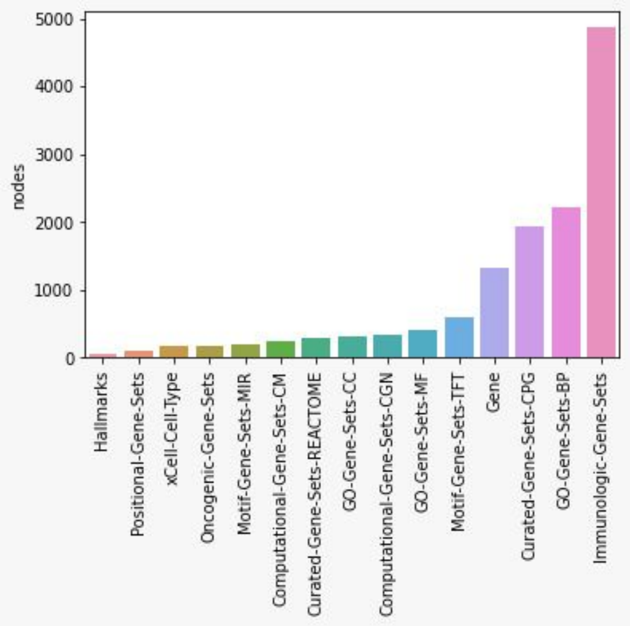
Types of samples in PCOS dataset

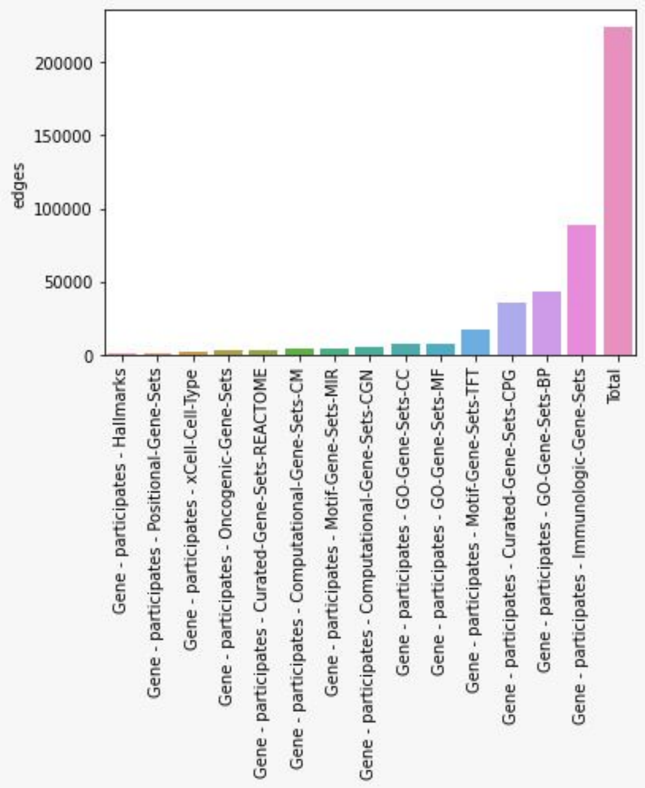
Total 175 samples from 9 different datasets; 1671 gene expressions

# Datasets: Initial Biological Inferences

Heterogeneous network analysis on merged PCOS dataset*

*Not all IDs are mapped





Most genes that play a role in PCOS are largely related to **immunological** datasets!

# Techniques Used

1. PCA, ICA, NMF initial results

2. DAE, VAE parameters and reconstruction costs

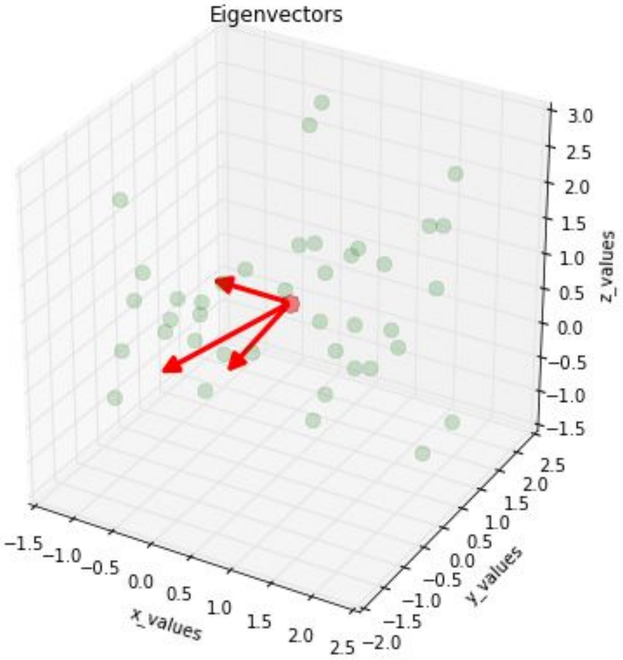
3. Gene set coverage, stability and SVCCA

4. Biological interpretations

** Way, Gregory P., Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, and Casey S. Greene. "Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations." Genome Biology 21, no. 1 (2020): 1-27.*

# Techniques Used

## PCA



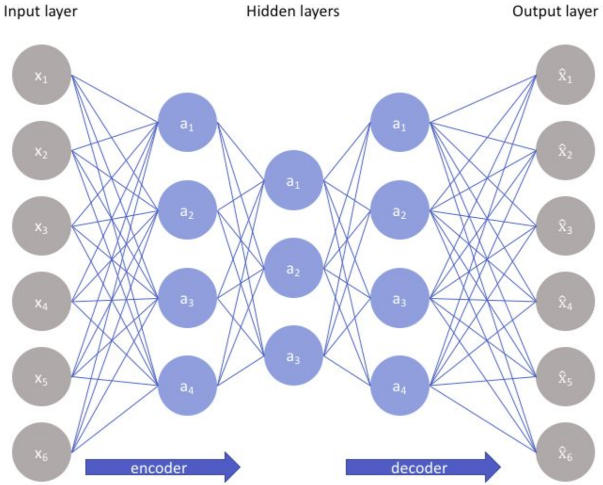
## ICA

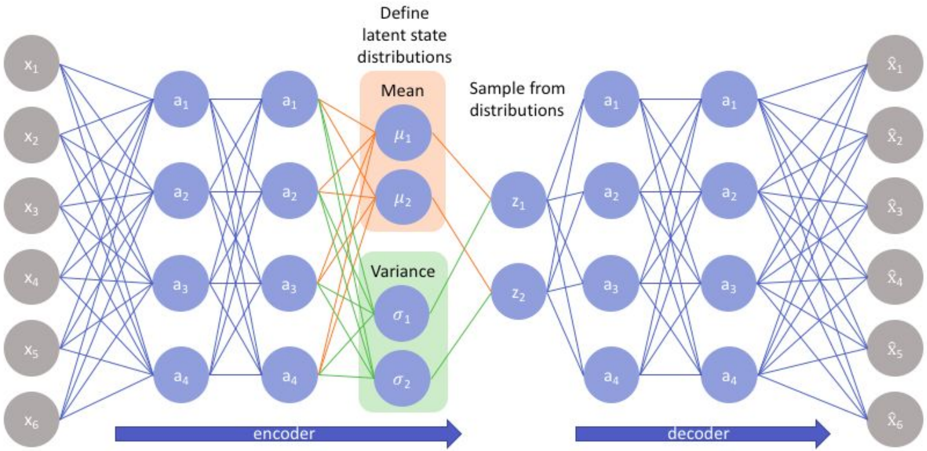
- The dimensions used are “independent to each other”
- Like a rotation of PCA

## NMF

- Like PCA, but, except the coefficients in the linear combination must be non-negative
- Dimensions that don’t contribute much have a zero coefficient.

# Techniques Used

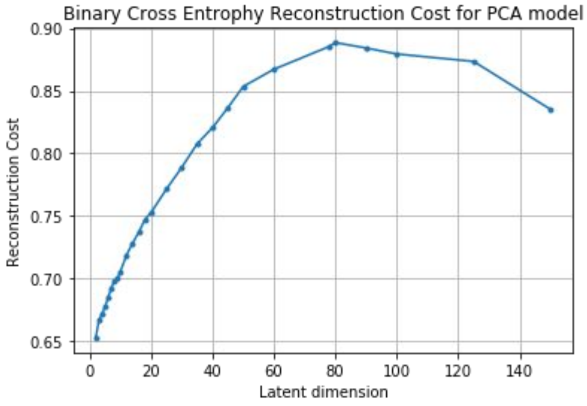


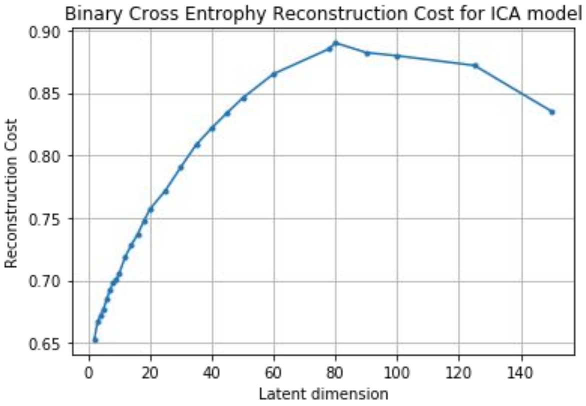


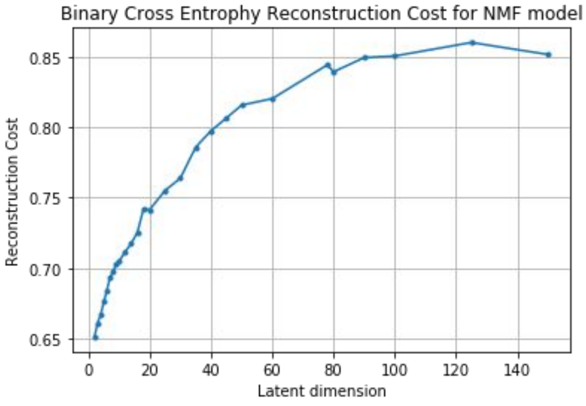
DAE models in addition, set a random fraction of the input data to 0.

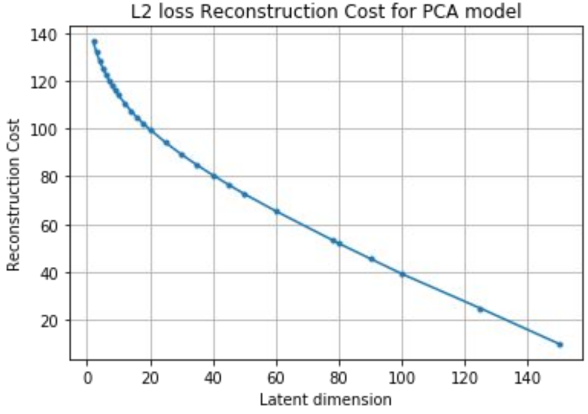
# Key results and interpretations

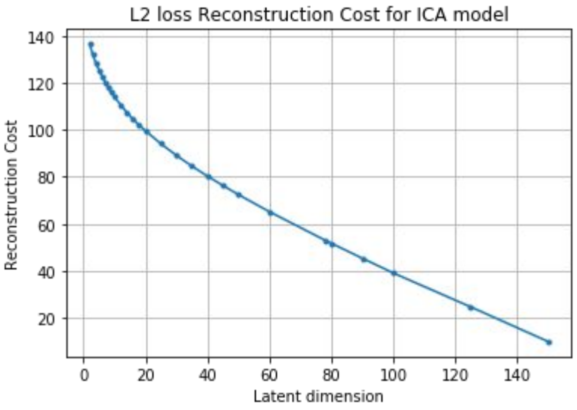
# PCA, ICA, NMF - Reconstruction Cost

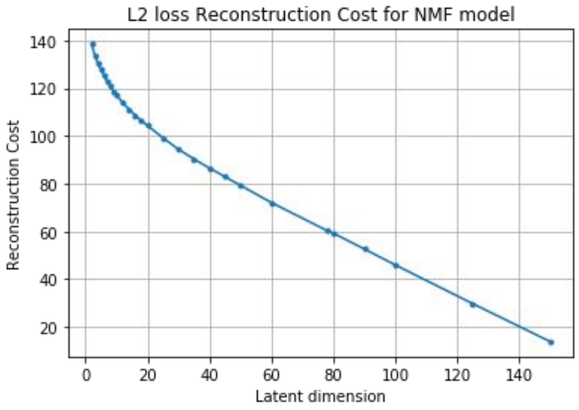

Binary Cross Entropy Reconstruction Cost for PCA model


Binary Cross Entropy Reconstruction Cost for ICA model

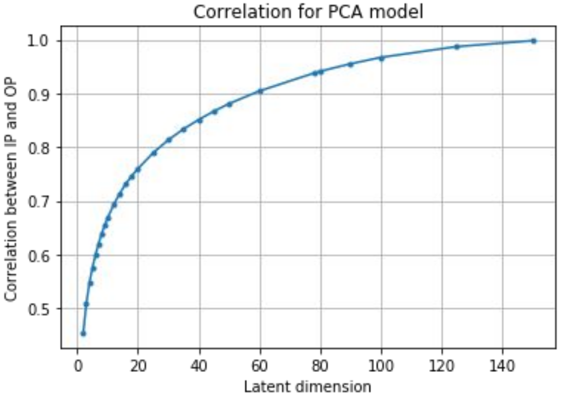

Binary Cross Entropy Reconstruction Cost for NMF model

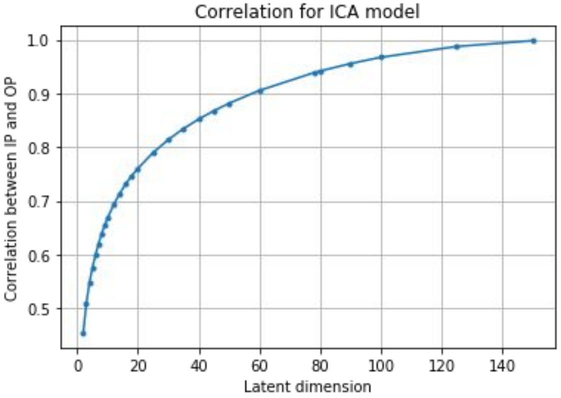

L2 loss Reconstruction Cost for PCA model

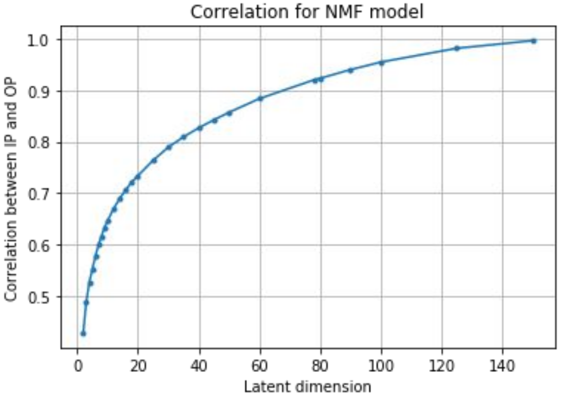

L2 loss Reconstruction Cost for ICA model

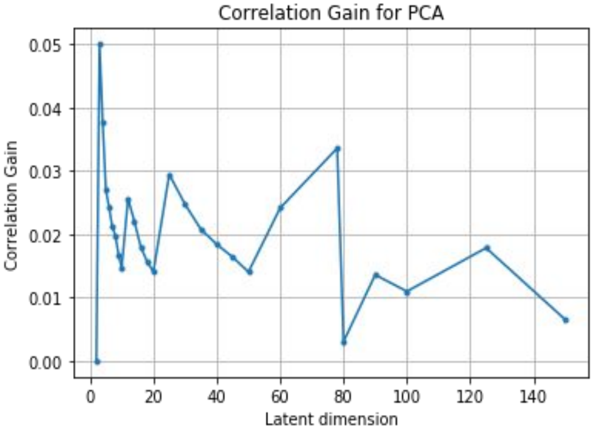

L2 loss Reconstruction Cost for NMF model

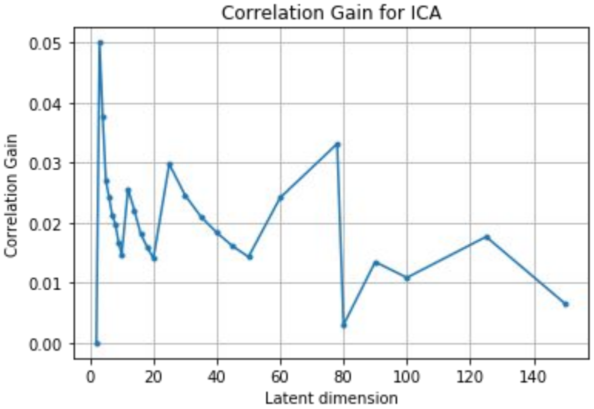
# PCA, ICA, NMF - Strongly Associated Dimensions

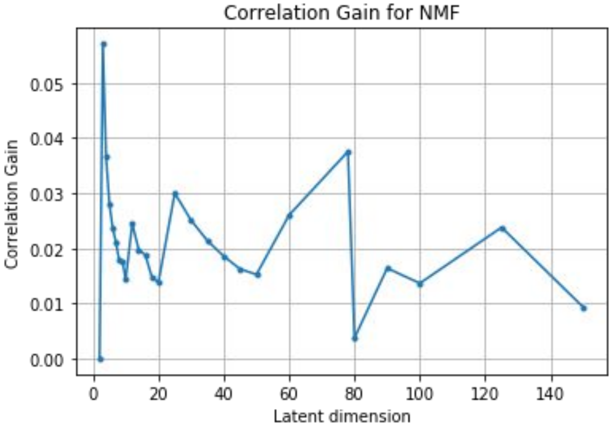

Correlation for PCA model


Correlation for ICA model


Correlation for NMF model

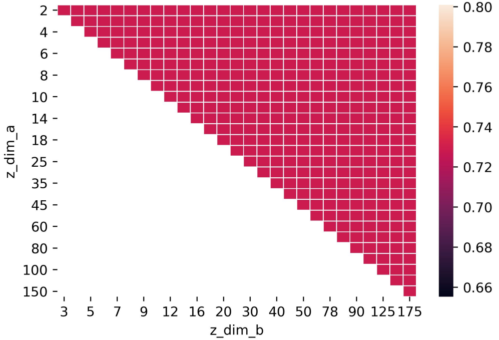

Correlation Gain for PCA


Correlation Gain for ICA

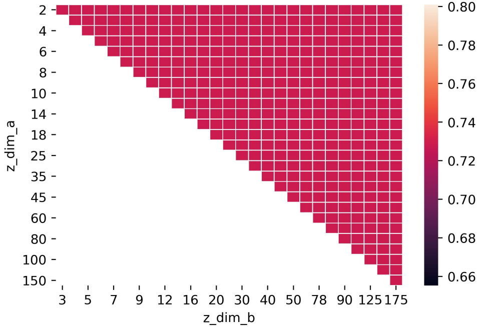

Correlation Gain for NMF

# PCA, ICA, NMF - SVCCA

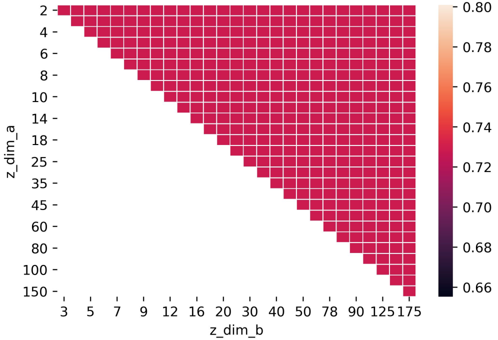
PCA: SVCCA means across latent dimensions (BCE loss)



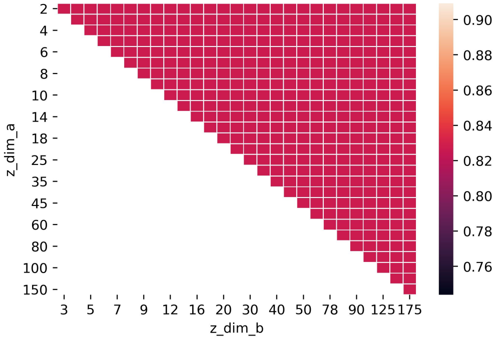
ICA: SVCCA means across latent dimensions (BCE loss)



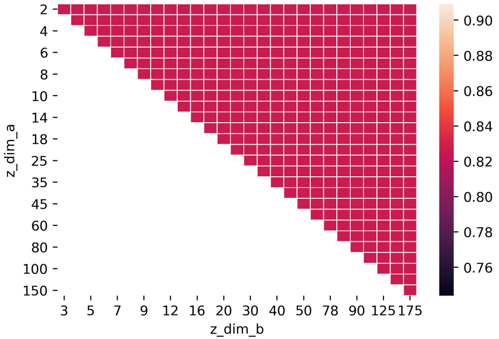
NMF: SVCCA means across latent dimensions (BCE loss)



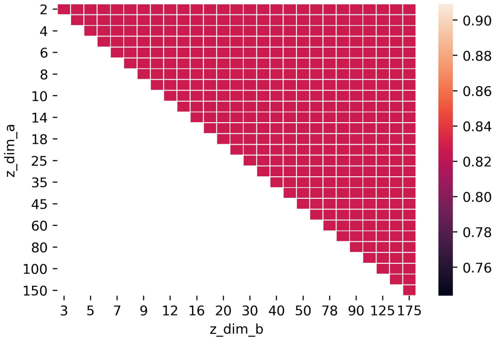
PCA: SVCCA means across latent dimensions (L2 error)



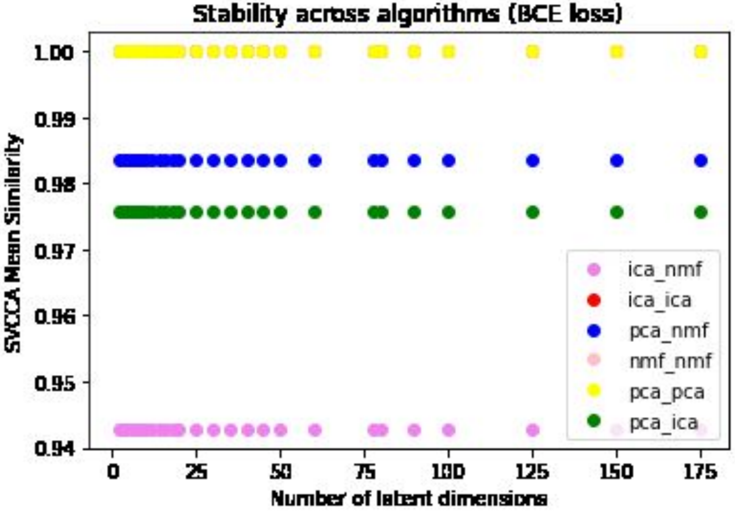
ICA: SVCCA means across latent dimensions (L2 error)

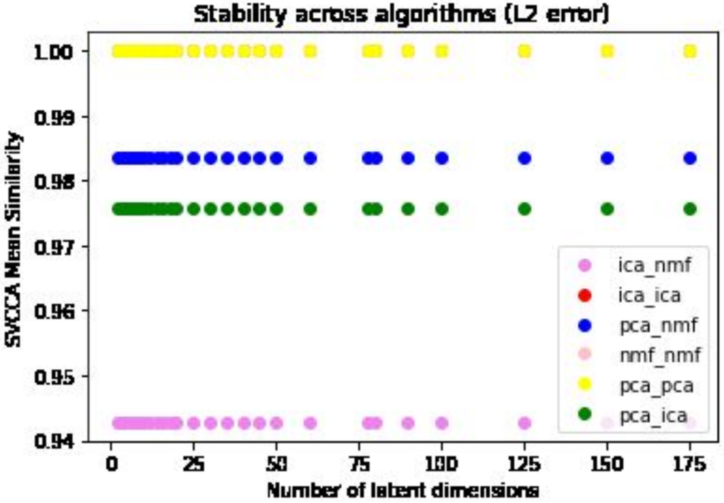


NMF: SVCCA means across latent dimensions (L2 error)

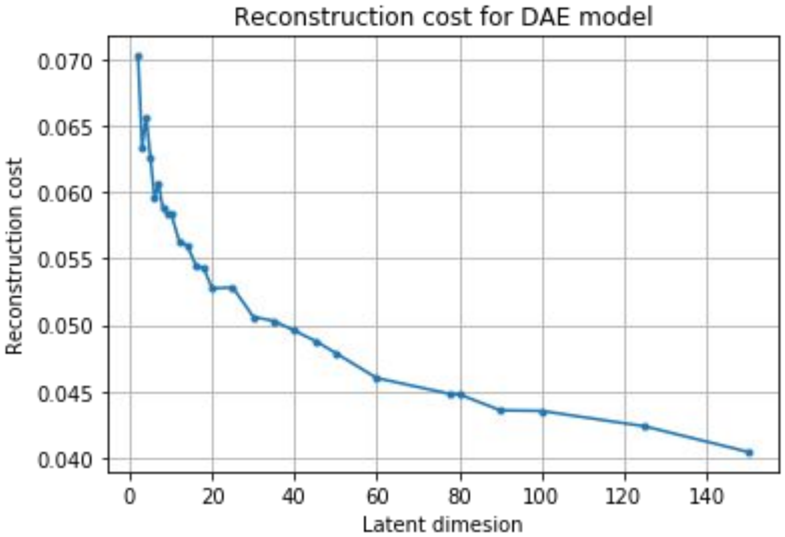


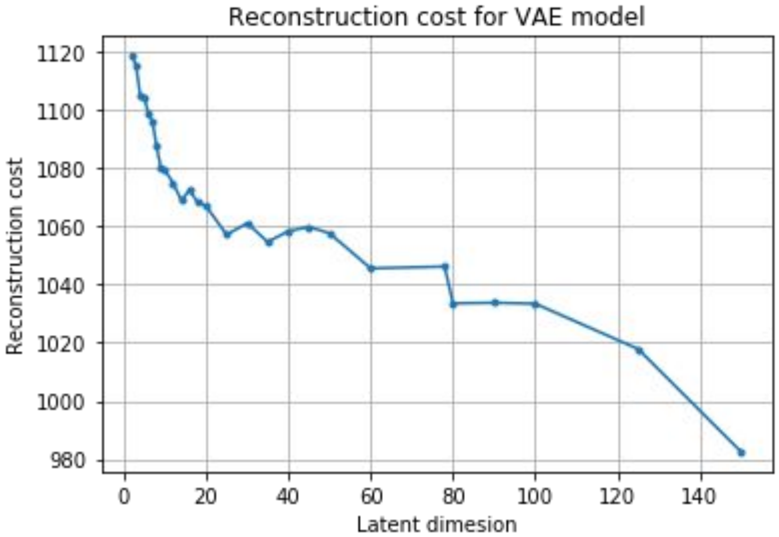
# PCA, ICA, NMF Correlation and Stability



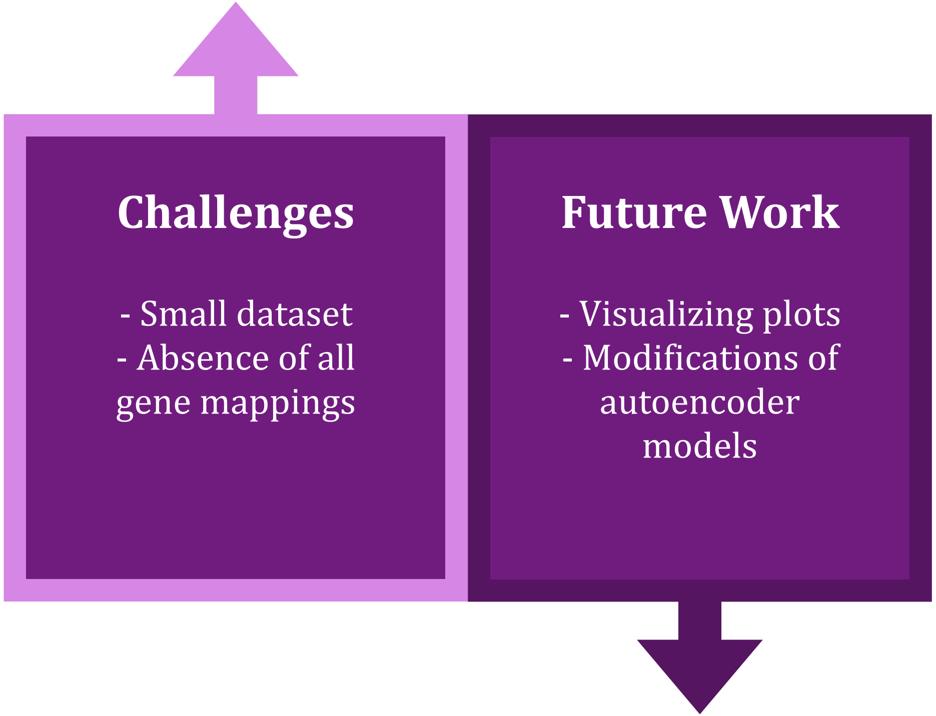


# DAE, VAE - Reconstruction Cost





# Challenges and Future Work



## Challenges

- Small dataset
- Absence of all gene mappings

## Future Work

- Visualizing plots
- Modifications of autoencoder models

**Thank you!**
**Any questions?**