Leveraging VAE to Dissect Inter-individual Variation

Ryan Tong[1], Dr. Lei Hou[2]

[1]Basis Independent Brooklyn, 556 Columbia St, Brooklyn, NY 11231; [2]Boston University, Department of Medicine, Biomedical Genetics Section 72 E Concord St. E244A, Boston, MA 02118

Introduction

Background:

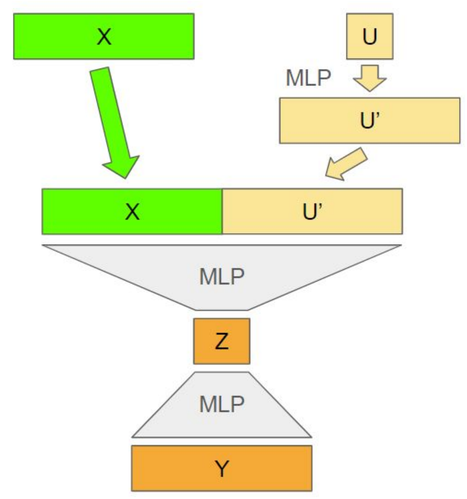
- Brain-related diseases are affected by genetic (G) and environmental (E), which can both be reflected in gene expression
- We aim to identify the latent factors associated with inter-individual variation in gene expression data
- There are, however, many different latent factor analysis frameworks, principal component analysis (PCA), independent component analysis (ICA), variational autoencoders (VAE), etc.

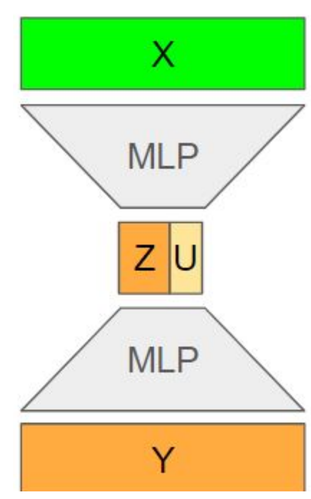
This Study:

- This study focuses on the capabilities of VAE to find latent factors
- We hypothesize that adding prior knowledge, e.g., age, sex, and smoking history, will improve VAE performance, resulting in the identification of biological factors related to brain diseases
- We look into flexible frameworks, such as identifiable VAE (iVAE), that allow the integration of prior knowledge

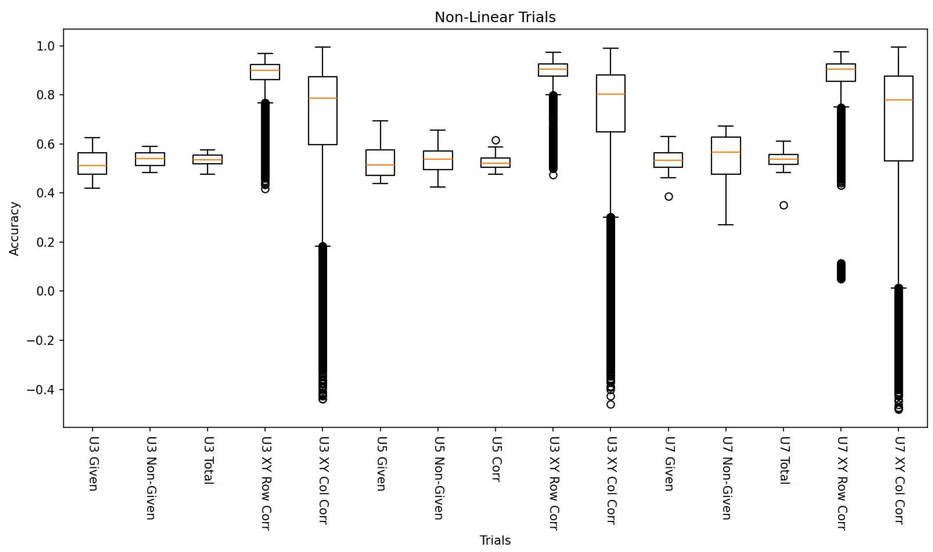
Methodology

- For now, research has mainly been done on simulated data, where the latent factors are already known
- The simulated data was fed through different VAE frameworks that incorporated prior knowledge
- 20 Trials were done for each framework
- Accuracies of latent layer (Z) predictions were compared to the known latent factors
- Additionally, row and column accuracies of the result of the VAE (Y) were compared to the known input data set (X)



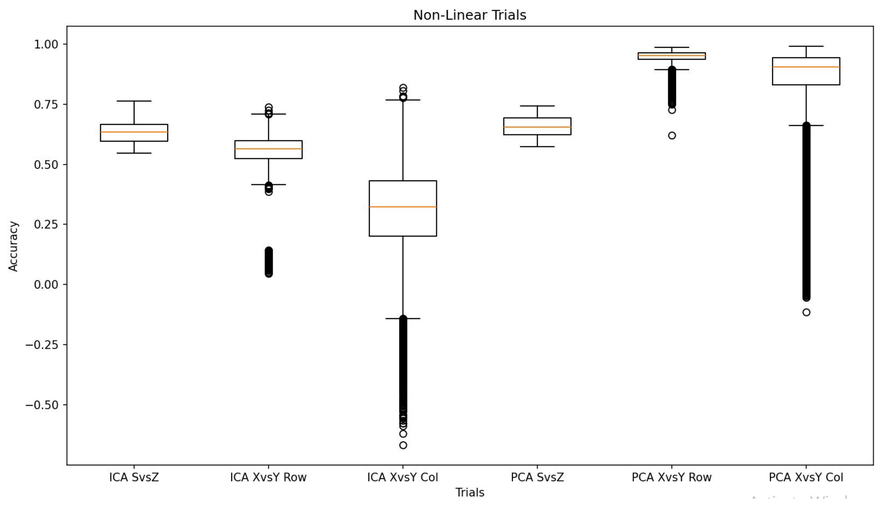


Results



VAE Trials

ICA and PCA Trials



Discussion

- The encoders of the VAE models we tested were found to have similar results to PCA
- But the encoders performed significantly better than ICA
- Unexpectedly, we found giving more prior knowledge did not result in better performance
- This could be due to the simplicity of our simulation data, which was generated through a 3-layer neural network

Future Directions

Further research will be conducted on more complex data to determine the effectiveness of our VAE networks in identifying inter-individual variation in real biological scenarios. Complex data includes simulated data from deeper neural networks and real biological gene expression data.

Acknowledgements

I would like to express my sincere gratitude to Professor Hou and the members of his lab for providing me with the invaluable opportunity to expand my computational biology knowledge and providing me with my first full time lab experience. Additionally, I would like to extend my gratitude to Boston University for giving me a chance to work with the shared computing cluster.