Developing a deep neural network model capable of conducting virtual gene manipulation experiment

Minh Nhat Nguyen, Son Le, David Tran

Dept of X, Bridge Institute, University of Southern California, Los Angeles, CA , USA

Bridge UnderGrad Science (BUGS) ***Summer Research Program***

USC University of Southern California

Abstract

The common approach of determining master regulator genes (MR) in disease has relied largely on trial-and-error experimental manipulation of individual genes, followed measuring the impact on disease phenotype. However this method is inefficient and incapable of identifying cooperative MR specific to individual patients. Therefore, there is a great need for new algorithm that can identify cooperative MR specific to individual patients by virtual manipulation in computational simulation.

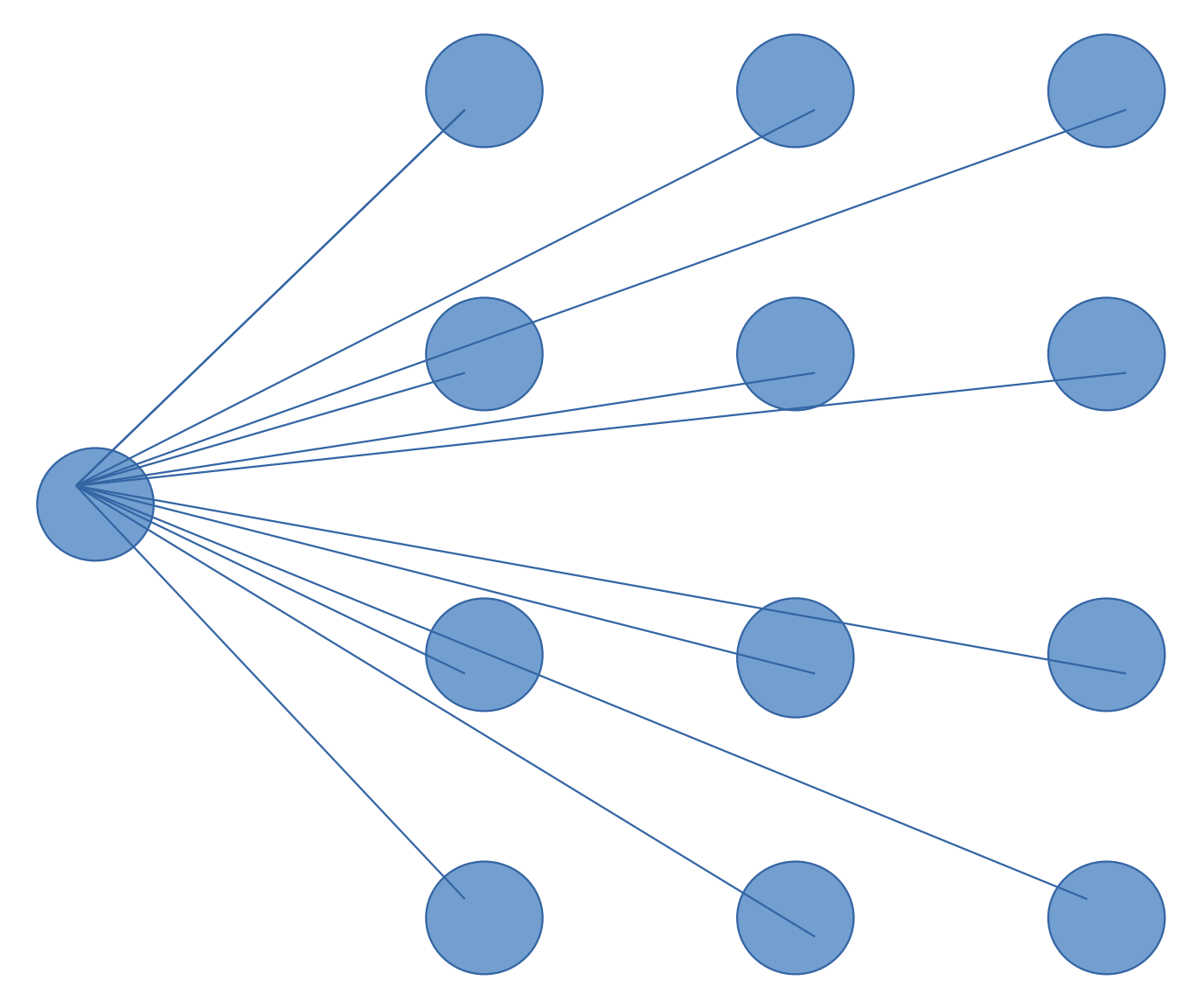
Auto-encoder model is an effective tool to model complex unknown relationships in biology due to their ability to learn and capture meaningful features and patterns. Utilizing transfer learning, we could train an auto-encoder network model based on large scale public data, and then fine tune it to a specific downstream task, identifying candidate therapeutic targets for specific disease. This approach can help accelerate discovery of key network regulators and candidate therapeutic targets.

Objectives

- Design an auto-encoder model capable of mapping genes-genes connection within specific human cell, with the model being trained on 237,824 training data points from RNA-seq dataset from Gene Expression Omnibus (GEO)

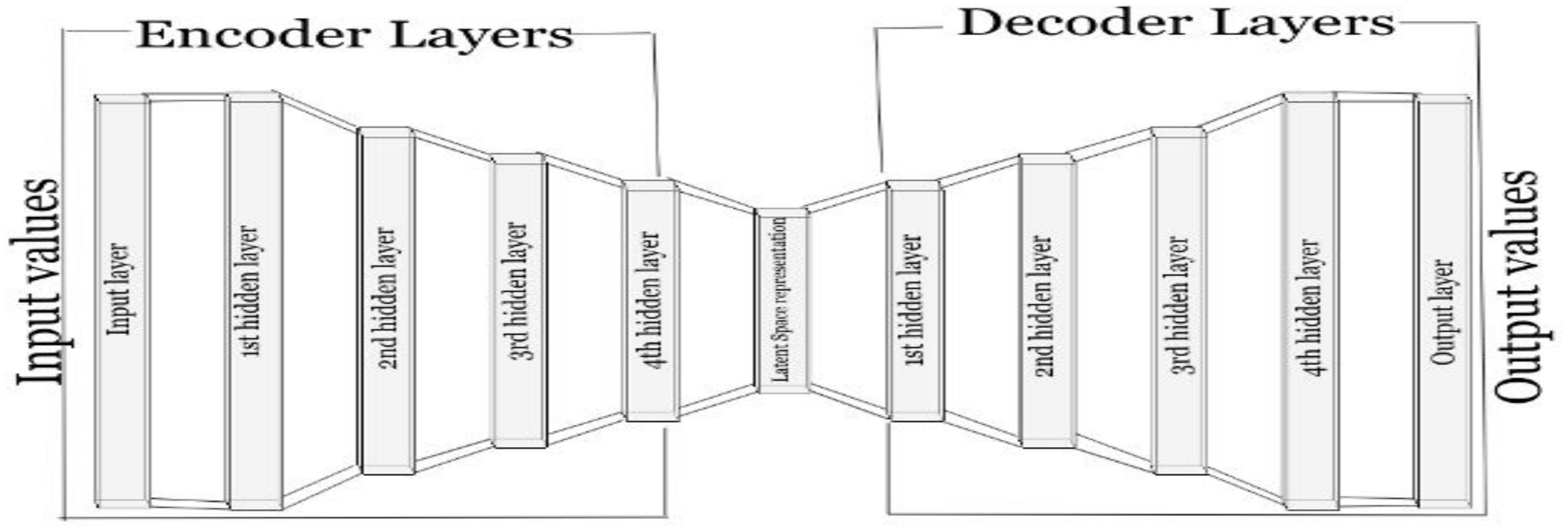
-Assess the performance of the model by judging it's ability to model the gene gene relationships by comparing the model predicted values with the ground truth.

HPC's Nodes Configurations



-Type of GPU: a100
-Total GPUs: 96
-Memory per CPUs: 32GB

Cell State Model Architecture

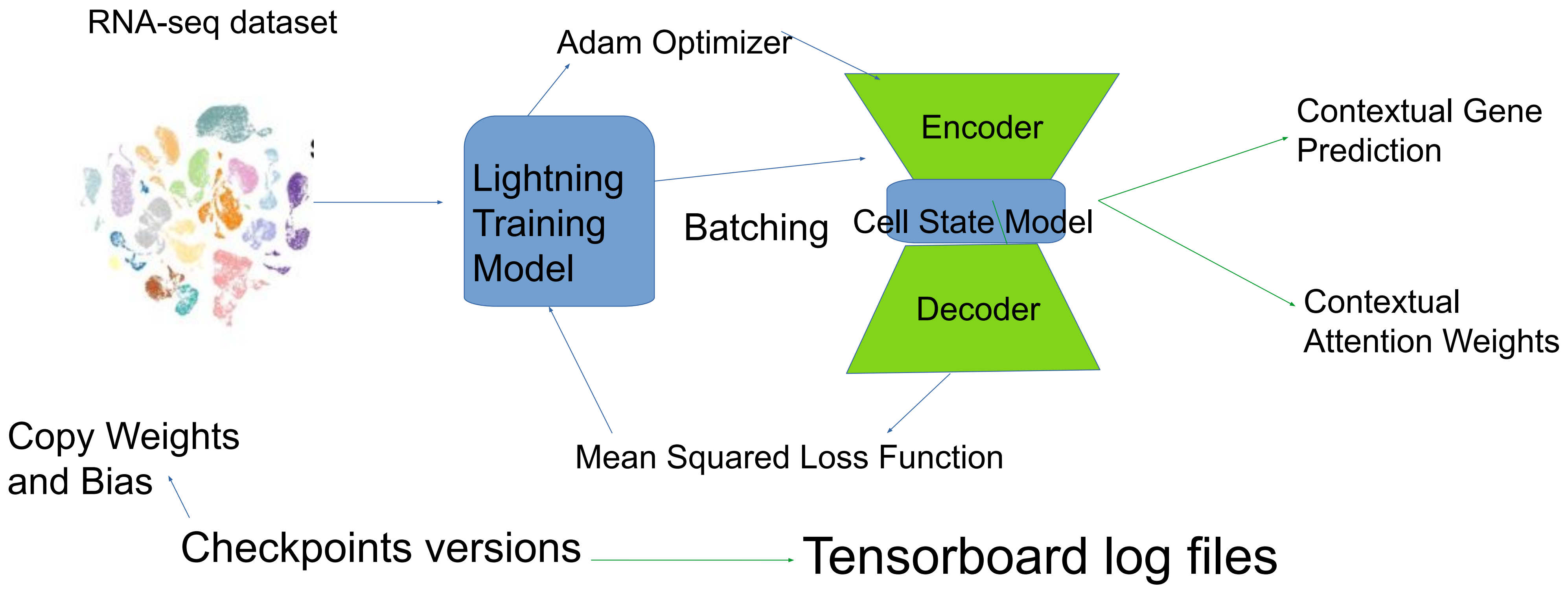


- Input layer's dimension: 19383 neurons
- Encoder Layers:
  + 1st layer: 20000 neurons
  + 2nd layer: 8000 neurons
  + 3rd layer: 4000 neurons
  + 4th layer: 2000 neurons

-Latent Space's Dimension: 1000 neurons

- Output layer's dimension: 19383 neurons
- Decoder Layers:
  + 1st layer: 2000 neurons
  + 2nd layer: 4000 neurons
  + 3rd layer: 8000 neurons
  + 4th layer: 20000 neurons

Model's self supervised training process



RNA-seq dataset

Adam Optimizer

Lightning Training Model

Batching

Encoder

Cell State Model

Decoder

Contextual Gene Prediction

Contextual Attention Weights

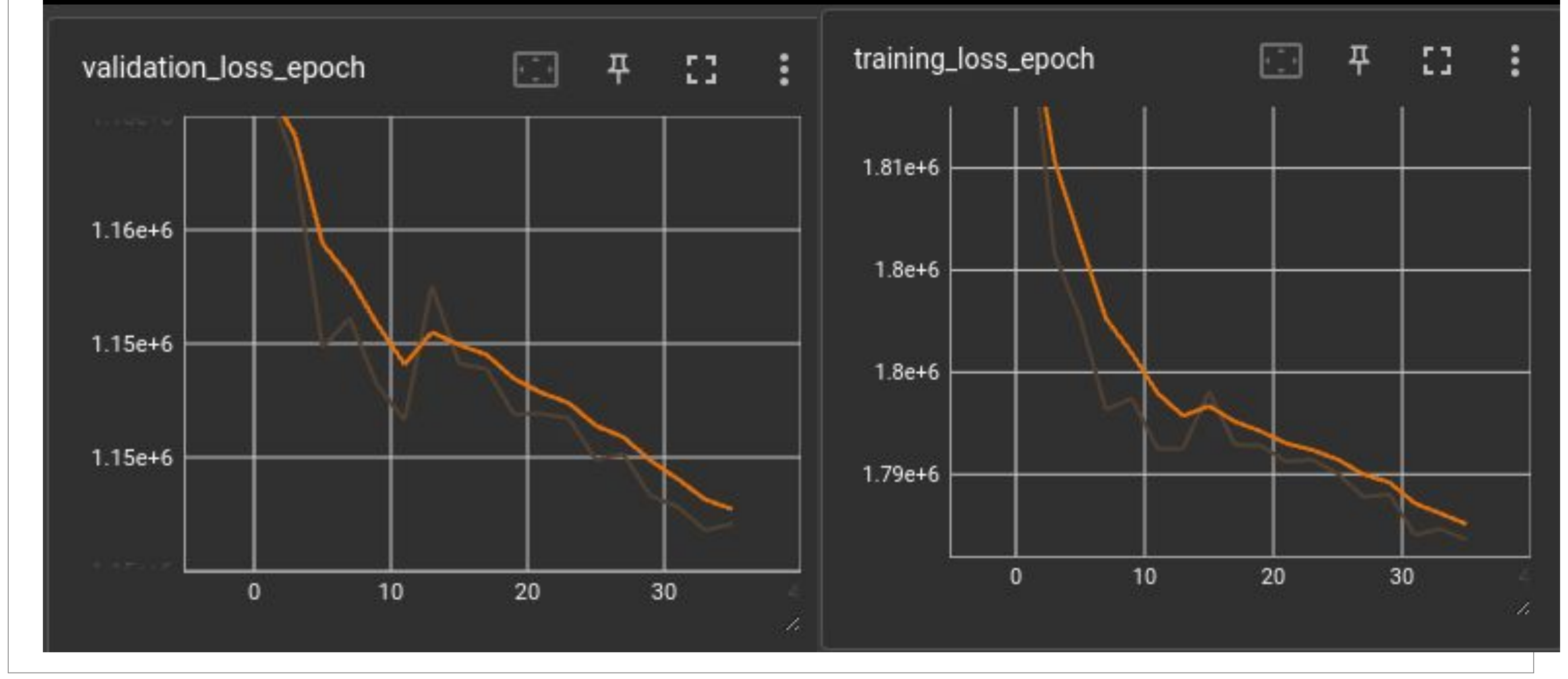
Copy Weights and Bias

Mean Squared Loss Function

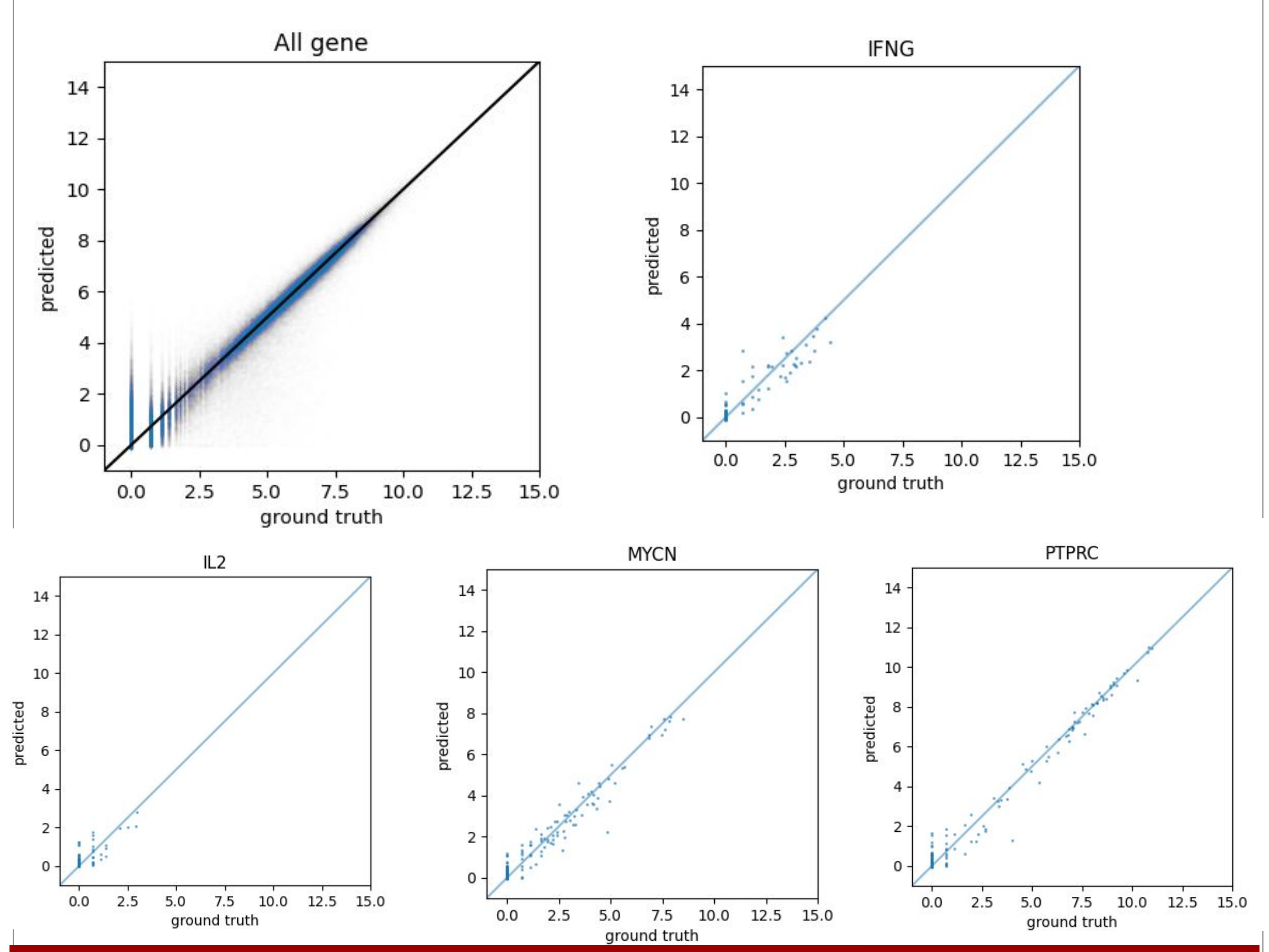
Checkpoints versions

Tensorboard log files

Model's Training Progress: Mean Squared Loss vs Epoch



Model Training Result



Assessment

-Model was able to reduce training loss and validation loss significantly after 35 epochs and with coefficient of determination (R square) = 0.94

-The graph shows the network perform poorly when the both the expected and predicted genes value are low with only around 30% accuracy for values around 0-2.5 tpm region

-The graph show that the network performance and accuracy increase as the value of predicted and ground truth became larger. This is due to error has less effect on high value comparing to low

Summary

Utilizing the auto-encoder model, the network were able to recognize and learn gene gene interactions after being trained with coefficient of determination 0.94

References: [1] Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in Network Biology. Nature, 618(7965), 616–624. https://doi.org/10.1038/s41586-023-06139-9

CONTACT US

bridge.usc.edu/bugs
Email:nn_990@usc.edu