Using Convolutional Autoencoders to Predict the Effects of Mutations on Protein Stability

Aum Khatlawala - 2020113008
Sarthak Aggarwal - 2020101008

April 17, 2023

1 Introduction

Proteins are essential biomolecules that perform a wide range of biological functions in living organisms. For a protein to function properly, it must be able to maintain its three-dimensional structure under a variety of circumstances. Protein stability can be significantly impacted by mutations, which are modifications to a protein's DNA sequence. Understanding the effects of mutations on protein stability is a challenging topic with significant consequences in areas like drug discovery, protein engineering, and personalised medicine to comprehend and forecast the effects of mutations on protein stability.

In this study, we propose to predict the impact of mutations on protein stability using Convolutional Autoencoders (CAEs) in conjunction with Fully Connected Neural Networks (FCNNs). An artificial neural network called a CAE may be trained to learn how to represent the structural data of proteins in a low-dimensional latent space. We can utilise the advantage of the encoding abilities of CAEs to capture the crucial characteristics of non-mutated proteins by training a dataset of wild-type proteins with known stability values.

The proposed method involves training a CAE to produce latent space representations of both wild-type and mutant proteins. Then, an FCNN framework that has been trained to anticipate the change in stability caused by the mutation will be fed the difference between the latent space vectors of the wild-type and mutant proteins. Once trained, the FCNN can be used to predict how changes in protein stability will affect previously unknown mutant proteins.

The relationship between protein mutations and stability may be better understood through the use of this research, which also has potential applications in the fields of protein engineering, drug development, and personalised medicine. The methodology, results, and discussion of our proposed approach's implications and limitations are covered in more detail in the following sections.

2 Baseline Model

Name: ProS-GNN: Predicting effects of mutations on protein stability using graph neural networks.

Authors: Shuyu Wang, Hongzhou Tang, Peng Shan and Lei Zuo.

Code Availability: ProS-GNN on Shuyu Wang's GitHub

Model Used: Combination of Gated GNNs and an MLP.

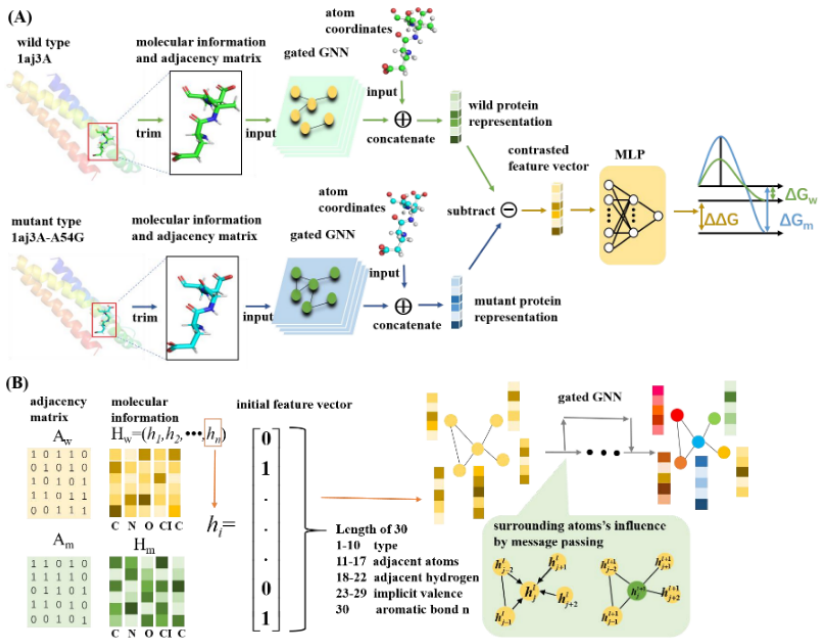

Figure 1: ProS-GNN Architecture and Certain Intricacies

Dataset Used: The dataset used in the stated baseline is the S2648 which contains both the PDB files of wild protein as well as mutant protein along with the $\Delta\Delta G$ values which is a parameter to measure the change in stability of a protein structure. The dataset is a subset obtained from the ProTherm dataset.

Running the Baseline Model: We were able to compile and run the baseline model of ProS-GNN on our local systems and observed that the model outperforms the previous work (ThermoNet) by a considerable margin.

The train-test data split was a 90-10 random split in terms of percentages. The total number of data points was 2160, each being a pair of the mutated and the wild-type protein.

We trained the model for 300 epochs and observed that the time taken for each epoch was around 33 seconds.

Results of the Baseline Model: The training loss curve obtained was as follows:

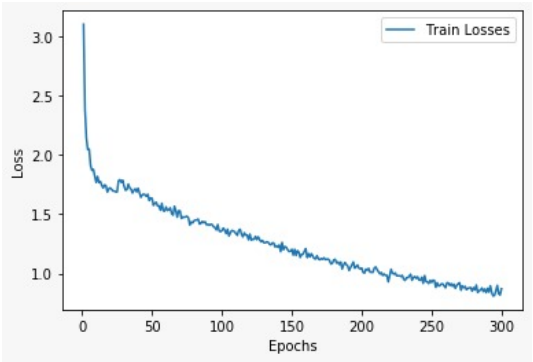

Figure 2: Training Loss Curve

We can observe that the training loss keeps decreasing after every epoch indicating the model is training effectively. The training loss reaches 0.870 after 300 epochs.

The Pearson Correlation Coefficient between the predicted and actual output obtained for the test set was as follows:

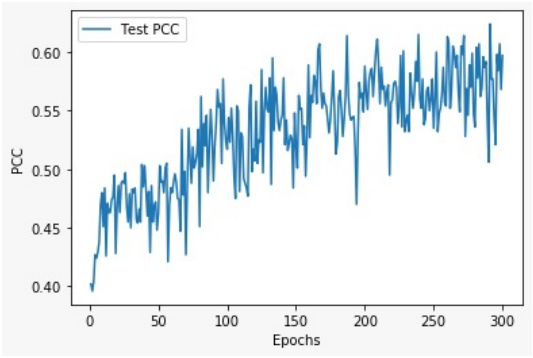

Figure 3: PCC Curve for Test Set

We can observe that even though the variance in the PCC test curve is quite high, the general trend is that it increases from 0.40 at the first epoch to around 0.60 after 300 epochs. This shows that the model is able to make significantly better predictions as and when it trains more. The ideal value for PCC is 1 and this is achieved when the predicted value and the actual value match for all the datapoints in the test set.

Comparing the Results Obtained:

Parameter	RMSE (test)	PCC (test)
ProS-GNN (Claimed)	1.23	0.61
ProS-GNN (Running locally)	1.15	0.597
ThermoNet (Baseline for ProS-GNN)	1.56	0.47

We can clearly see that ProS-GNN is able to perform at the level they have claimed in the paper. Along with this, it is able to outperform the previous state-of-the-art model that they used as baseline (ThermoNet) by a considerable margininj terms of the two parameters that have been used in by all the previous papers performing this task, namely:

1. RMSE - Ideal value is 0. The previous best was 1.56 and the baseline achieved 1.15.
2. PCC - Ideal value is 1. The previous best was 0.47 and the baseline achieved 0.6.