## Student's Declaration

I hereby declare that the work presented in the report entitled **"Classification of Agonist and Antagonist using Machine Learning"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Biosciences* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Arjun Ray**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

**Bhavay Aggarwal** **Place & Date: IIITD 10/5/21**

## Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

**Dr. Arjun Ray** **Place & Date: IIID 10/5/21**

**Abstract**

Proteins are essential for carrying out the activities our cells perform. Agonist and Antagonist molecules are essential in dictating cell activities and hence the Identification of the action of such complexes will help enhance drug discovery and production. We create a dataset of such proteins and then apply machine learning techniques to classify them.

Keywords: Proteins, Agonists, Antagonists, Machine Learning, Classification

## Acknowledgments

I would like to express my sincere gratitude to our advisor Dr.Arjun Ray for the continuous support in my study and related research, for his patience, motivation, and immense knowledge. His guidance helped us through out the research so far and in writing of this report.

# Contents

# Chapter 1

# Introduction

Proteins are essential for carrying out the activities our cells perform. An agonist molecule when bound to a receptor produces a response within the cell hence the complex is termed as agonist complex. An antagonist will block the binding site/produce the opposite effect of that of an agonist. An example of this can be opioid receptor antagonists which are well known for their ability to attenuate or reverse the effects of opioid agonists. This property has made them useful in mitigating opioid side effects, overdose and abuse. Identification of the action of such complexes will help enhance drug discovery and production.

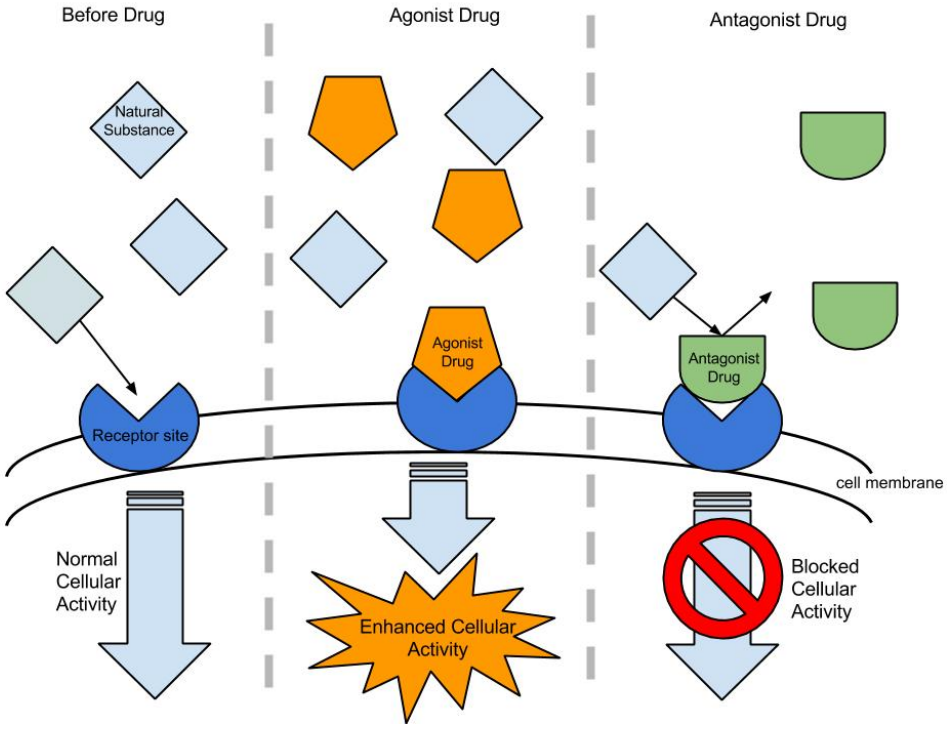



Figure 1.1: An image illustrating the action of agonists and antagonists

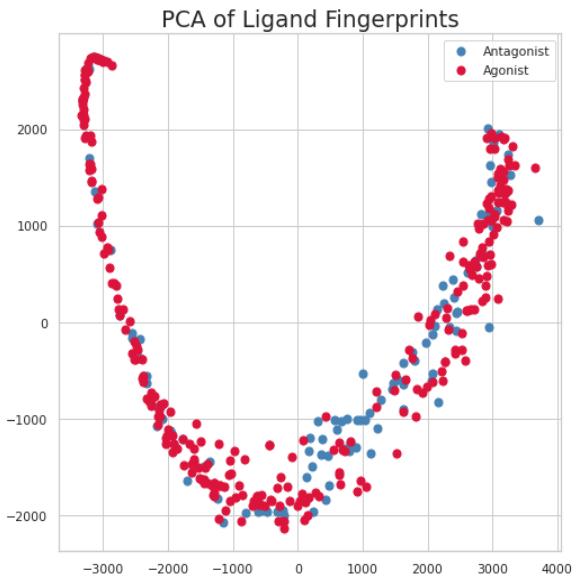
# Chapter 2

# Data Collection

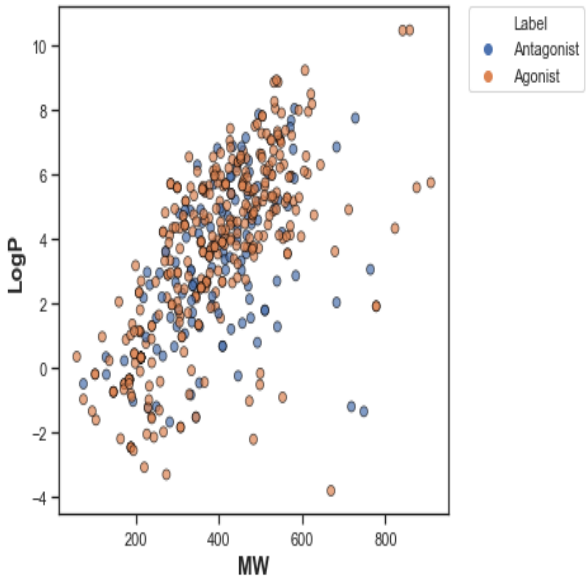
For collecting the relevant protein files, RCSB database was used. All ids matching with the search terms " Protease bound with agonist " and "Protease bound with antagonist " were extracted. Files containing RNA/DNA were removed. Next step of automated filtering was to remove files which did not contain any words similar to either "Agonist" or "Antagonist" in their PDB title, header or its primary PubMed citation. Files were then manually filtered using the above mentioned criteria and also, on the ligands and molecules present in the file to verify that the ligand and protein are present in the file. Duplicates were removed and the final data composition was as follows –

- **551 Proteins**
- 156 Antagonists
- 395 Agonists

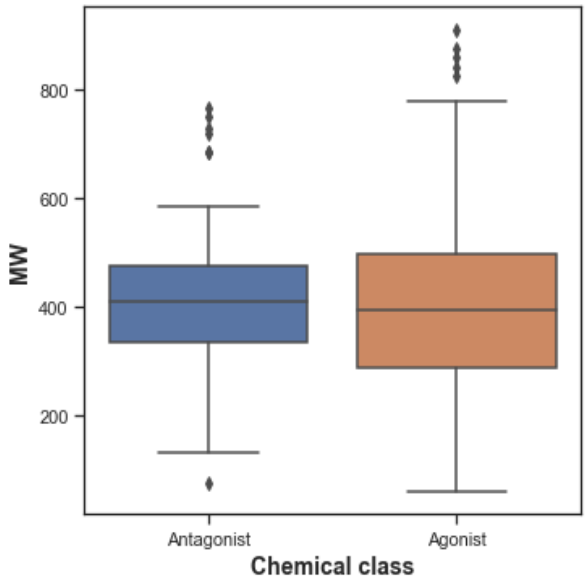
Peptides ligands were also removed when considering the ligands as features in the models which left 489 proteins in which 131 were antagonists and 358 agonists. Additionally, there are 214 unique proteins and 423 unique ligands.



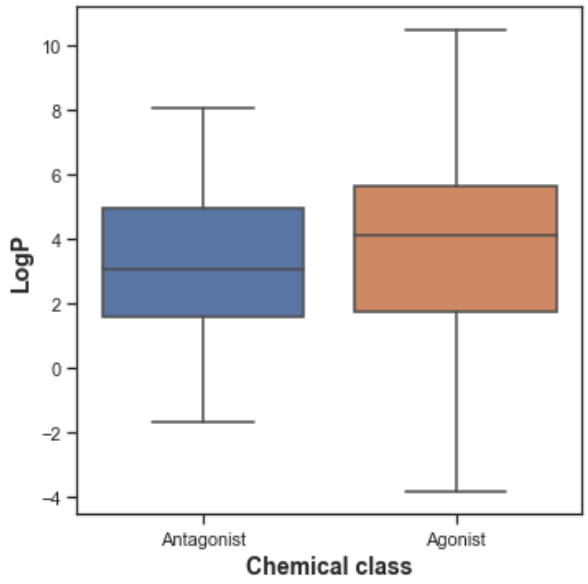

PCA of Ligand Fingerprints



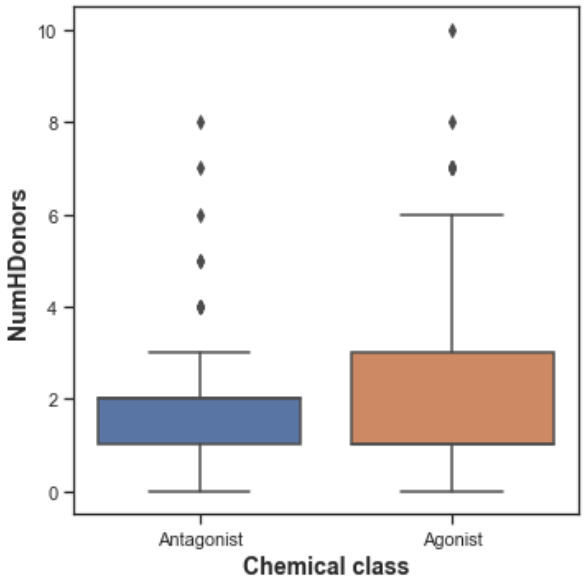

Molecular Weight vs Octanol-water partition coefficient



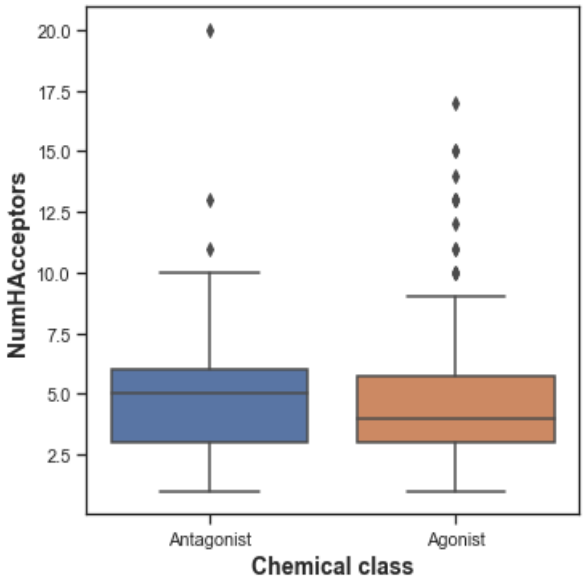

Molecular Weights of Ligands




Octanol-water partition coefficient of Ligands




Number of H-Donors of Ligands




Number of H-Acceptors of Ligands

Figure 2.1: EDA of Ligands

Above EDA shows no clear distinction between the 2 ligand classes - agonist and antagonist. Similar is the case with the proteins.

# Chapter 3

# Sequence Based Methods

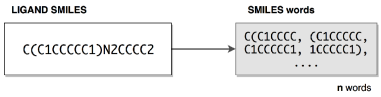
Protein Sequence is the chain of residues from which the protein is composed of. It contains no information regarding the 3d structure of the protein so one would not expect it to perform well but the results from ProtVec and advances and Natural Language processing make it a viable model.

ProtVec is a representation of the protein sequence as an overlapping sliding window of residues.

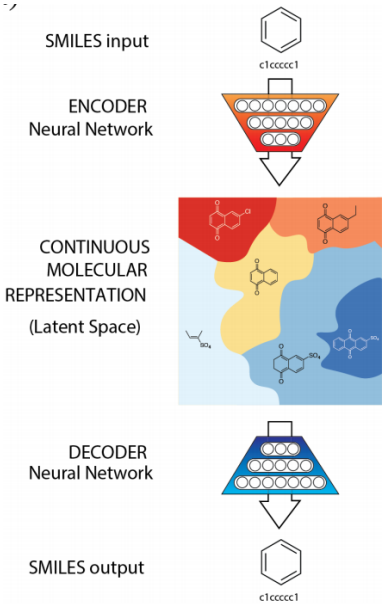

Figure 3.1: Dividing a protein into overlapping sliding windows with length k=3

Using Word2Vec, the sequences generated were embedded into vectors of length 100. The sum of these vectors is the ProtVec representation of the protein. The embeddings were added to create 1-100 length vector for each protein which were trained on 4 machine learning models - AdaBoost, XGBoost, Random Forest and SVM, with stratified 10 folds.

To incorporate the ligand, their SMILES was featurized by one-hot encoding, Word2Vec(SMILESVec) and latent space representation(ChemVAE). The result was concatenated with ProtVec features and trained on the ML models.




(a) SMILESVec




(b) ChemVAE

### Advanced NLP Methods

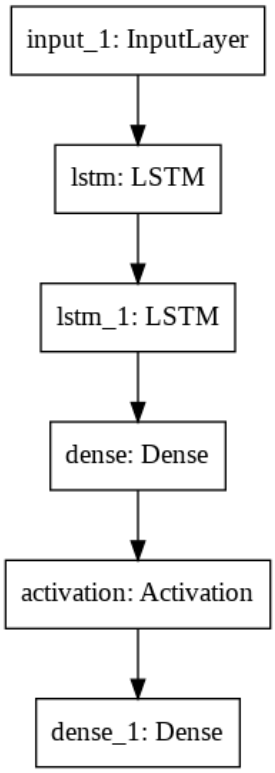
To experiment with advanced NLP we techniques, we first started with NBSVM - Naive Bayes Support Vector Machine which is a baseline language model. The preprocessing of protein sequences was the same as for ProtVec but ligands were not included here. Further we trained more advanced language models namely BERT, RoBERTa and ELECTRA. To further improve the performance of the models, the models were trained on 10k randomly chosen protein sequences and then fine-tuned on our dataset.

# Chapter 4

# Physiochemical Properties Based Methods

The physiochemical characteristics, such as molecular size, net charge and amino acid composition are some of the factors that contribute to the functional properties of proteins. These features are extracted solely from the sequence of the protein and again do not use the structural information present in a PDB file. 20 features were extracted from protein sequences using iFeature library. These features include –

- **Pseudo amino acid composition** – Protein sequence is characterized using matrix of amino-acid frequencies. Compared to AAC, additional information are also included such as correlation between residues of a certain distance.
- **Conjoint Triad** – Neighbor relation in protein sequence is calculated by encoding the sequence is continuous triads of residues.




The model used is described on the left. The double LSTM layers act as encoders. They are capable of learning the complex dynamics within the temporal ordering of input sequences as well as use an internal memory to remember or use information across long input sequences. Other models were trained using Bi-directional LSTM. Feature-set was expanded to include word2vec embeddings and residue index features from AAIndex. The model was hyperparameter optimized using Optuna.

# Chapter 5

# Graph Based Methods

Graph models help incorporate features such as accessible surface area which play a key role in protein binding. Residues can be represented as nodes and bonds as edges. Properties such as accessible surface area are embedded as node features. Distances between residues is also preserved.

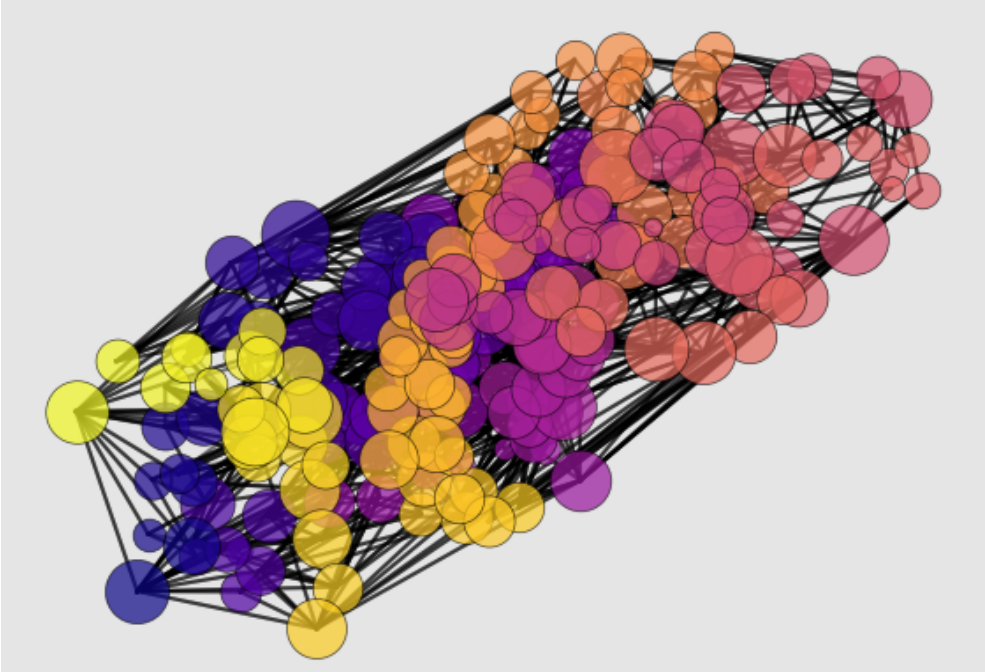


Figure 5.1: Graph generated of protein with id 6HLL

Graphs were generated using the python library Graphein. Node features selected were coordinates, ASA and RSA while edge features were based on distance and interactions. The following state-of-the art models were trained using DGL-LifeSci -

- Molecular graph convolution (Weave)
- Graph Convolutional Network (GCN)

- Graph Attention Network (GAT)
- Graph Attention Network (AttentiveFP)

GNN require a larger amount of data to make predictions, so we generated 10k graphs and trained the Graph Convolution Model. The prediction layers were then removed and the model was fine-tuned on our dataset. This was done for all 4 models with various hyperparameters being tried out

# Chapter 6

# Results

### Sequence Based Models

Models trained on ProtVec were able to achieve 80% accuracy and the addition of ligand features was able to bump it to 86%. AdaBoost was the best performing model in both cases.

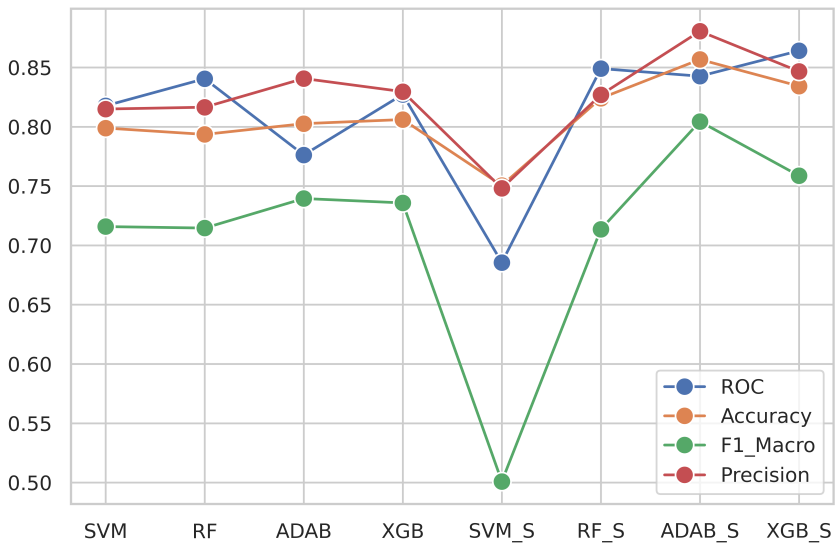



Figure 6.1: Results from sequence based models (w2v and w2v + smiles)

Out of 11500 features, only 47 were used by the AdaBoost model to make predictions. This suggests that by selecting these features, the model is able to get an advantage over SVM. SMILES representations contribute to 32% of the features used by AdaBoost and the boost in performance is a confirmation of our pairwise approach to the problem.

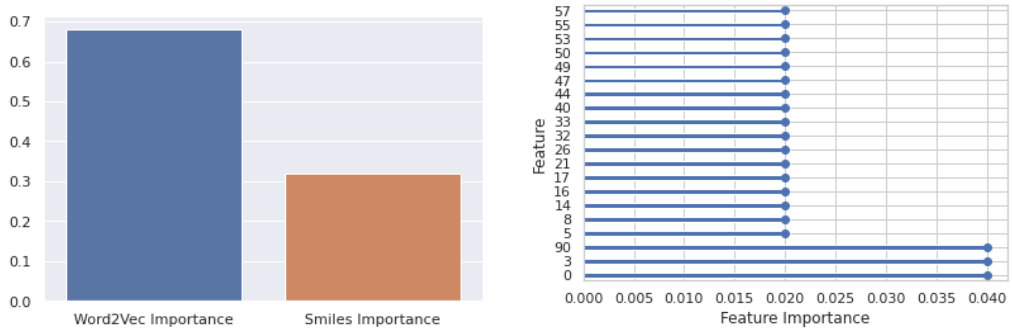



Figure 6.2: Feature Importance of AdaBoost

One-hot encoded representation of SMILES outperforms more efficient representations tested possibly owing to the our previous finding of only a few features being important. It would seem that information important to our classification task are lost in the dimension reduction, we also find that AdaBoost with one hot SMILES uses 47 (31 protein, 16 smiles) features whereas with Word2Vec embedded SMILES it only uses 42 (23 protein, 19 smiles) which is an unexpected decrease in the number of protein features being used by the model.

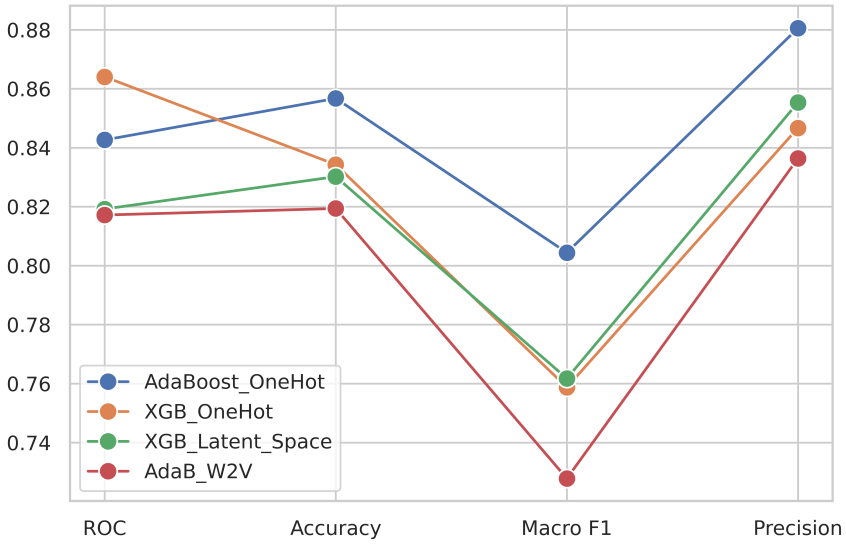



Figure 6.3: Results from SMILES representation experiments

### Physiochemical Properties Based Model

Using the 20 features extracted from IFeature and hyperparameter optimization using Optuna, we were able to achieve 80% accuracy. Addition of features namely AAIndex and ProtVec sequences led to a decrease in performance. Since, AAIndex is just a positional replacement of

amino acid with physicochemical feature values, they were not adding any value to our feature set. The performance decrease on the addition of ProtVec sequences can be attributed to the loss of sequence properties in the process of creating Word2Vec embeddings apart from the issues such as the size of the database which is not suitable for deep learning models. This approach had slightly better performance than vanilla Word2Vec but would require a much larger dataset to fully utilize the capabilities of LSTM's. The inclusion of ligand features might also boost the models performance.

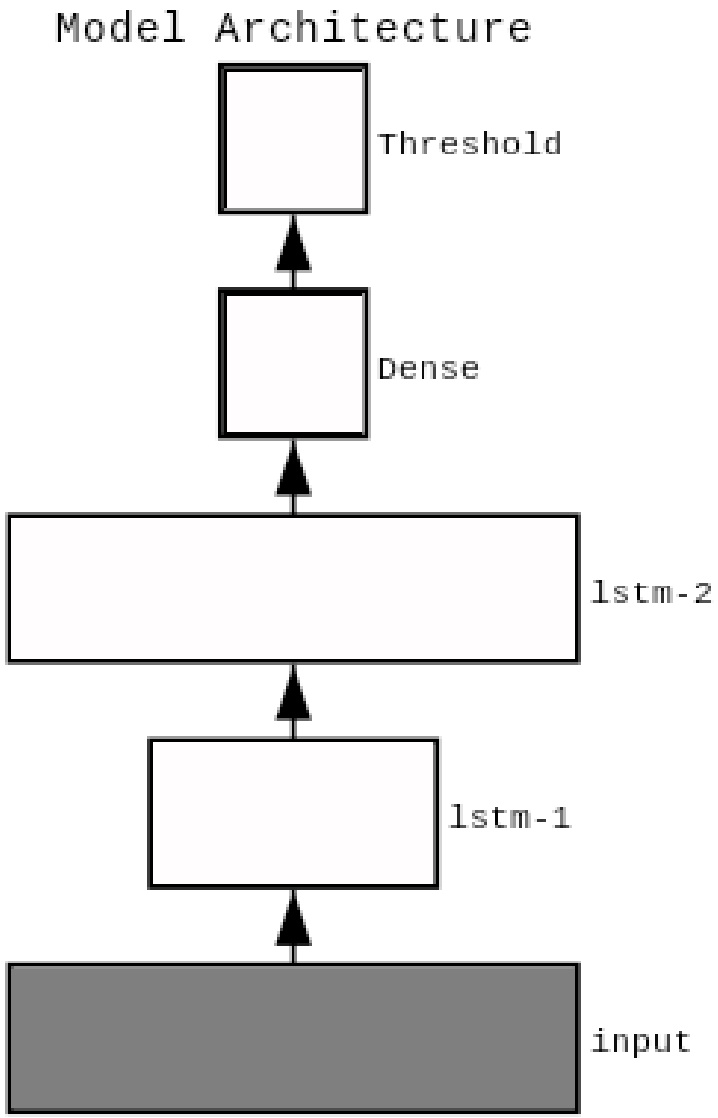



Figure 6.4: LSTM Model Architecture

**Graph models**

Graphs created by graphein were again trained with 10 stratified fold cross validation. All the four models have similar performance with GCN performing the best in terms of predicting the minority class i.e antagonists. Weave only predicted the majority class whereas GAT and AttentiveFP were able to predict antagonists but with poor precision. To overcome the training

issues caused by our relatively small dataset for such a task, the models were trained on 10k randomly chosen proteins. The classification layers were then removed from the trained model and the model was fine-tuned on our dataset. This had minimal impact on the performance of the models with GCN again performing the best but comparatively worse than other techniques.

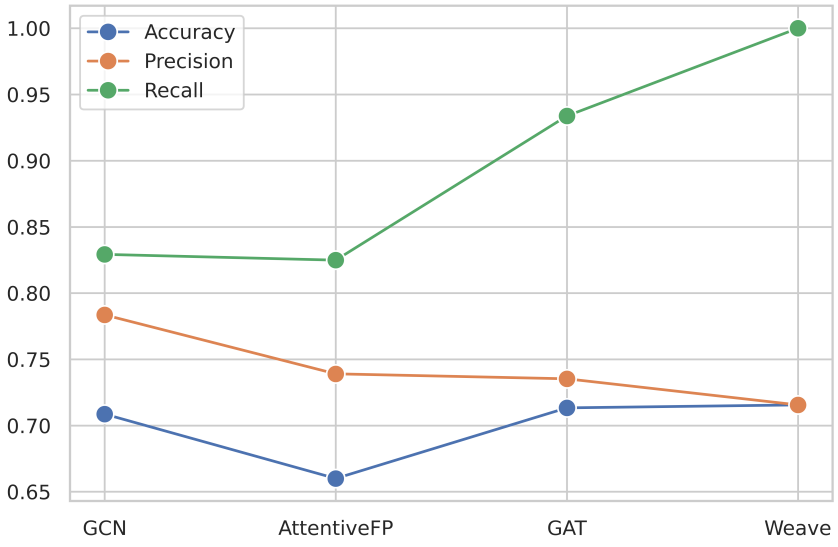



Figure 6.5: GNN Results

# Chapter 7

# Future Work

Entire graphs although give an accurate representation of protein structures, are harder to featurize to solve classification problems like ours. Also, Graph Convolutions do not necessarily suit problems like ours. Sequence Based Models show alot of promise especially if we can somehow incorporate structural features efficiently and generate accurate representations. Specialized Transformers can help achieve new levels of prediction performance with better generalization than standard ML models.