WiDS 2024
Drugweave - Secrets of the Sequence

Abhineet Majety

1 Overview and Dataset

This project aims to study the effectiveness of some neural network architectures on predicting drug-protein interaction. Obtaining reliable estimates of drug-protein interaction can have an impact on drug discovery as we can efficiently predict drug efficacy.

In this project, we have used open sourced Drug-Target Affinity Davis dataset. The dataset contains the affinity index (pK_d) of all possible pairs among 442 proteins and 62 drugs. Each protein is represented by its amino acid sequence and each drug is represented by its SMILES sequence.

2 Encoding Sequences

Various ways of encoding sequences were used:

- Mapping amino acids to integers from 0 to 20
- Mapping each symbol in SMILES sequences to one integer
- Encoding SMILES symbols using the python `rdkit` module

3 Train-Test splitting

Several ways of splitting data into train and test datasets were used:

- **No new proteins in test**: Each protein's interactions were split $70\% - 30\%$ between training and testing
- **New proteins in test**: 90% of the proteins were included in training with all drug interactions, while 10% of proteins were completely reserved for testing
- **No new drugs in test**: Ensured all drugs in testing were seen during training
- **New proteins in test**: Held out some drugs entirely for the test set

The models were evaluated on each of these train-test splits.

4 Architectures used

The data was trained on three different neural networks:

- **Feedforward Neural Network**: The network has fully connected layers with ReLU activation function.
- **Convolutional Neural Network**: Data augmentation of the train data was performed. The model uses convolutional, max-pooling and dense layers.

- **LSTM Model**: Protein and drug sequences are first passed through seperate LSTM networks to extract embeddings. The embeddings are then passed to a fully connected network.

5 Metrics

Four metrics were used to evaluate the models:

- Concordance index
- Mean square error
- Pearson correlation coefficient
- Area under precision recall curve (AUPRC)

6 Results

The following table shows the metrics for each model for all the train-test splits:

- No new proteins in test split

Model	Concordance Index	Mean Square Error	Pearson Correlation	AUPRC
Feedforward NN	0.63	0.91	0.21	0.13
CNN	0.67	0.76	0.30	0.18
LSTM	0.63	2.15	0.22	0.54

- New proteins in test split

Model	Concordance Index	Mean Square Error	Pearson Correlation	AUPRC
Feedforward NN	0.74	0.65	0.49	0.40
CNN	0.76	0.50	0.55	0.47
LSTM	0.76	1.70	0.55	0.55

- No new drugs in test split

Model	Concordance Index	Mean Square Error	Pearson Correlation	AUPRC
Feedforward NN	0.71	0.63	0.43	0.31
CNN	0.76	0.55	0.55	0.42
LSTM	0.72	2.16	0.42	0.54

- New drugs in test split

Model	Concordance Index	Mean Square Error	Pearson Correlation	AUPRC
Feedforward NN	0.68	0.76	0.28	0.03
CNN	0.77	0.45	0.43	0.06
LSTM	0.32	2.48	-0.16	0.55