## Final Project

# DeepPurpose: a Deep Learning Library for Drug-Target Interaction Prediction

Piyar Ali , Ali


CSCI E-104 Advanced Deep Learning, 2023

**Harvard University Extension School**

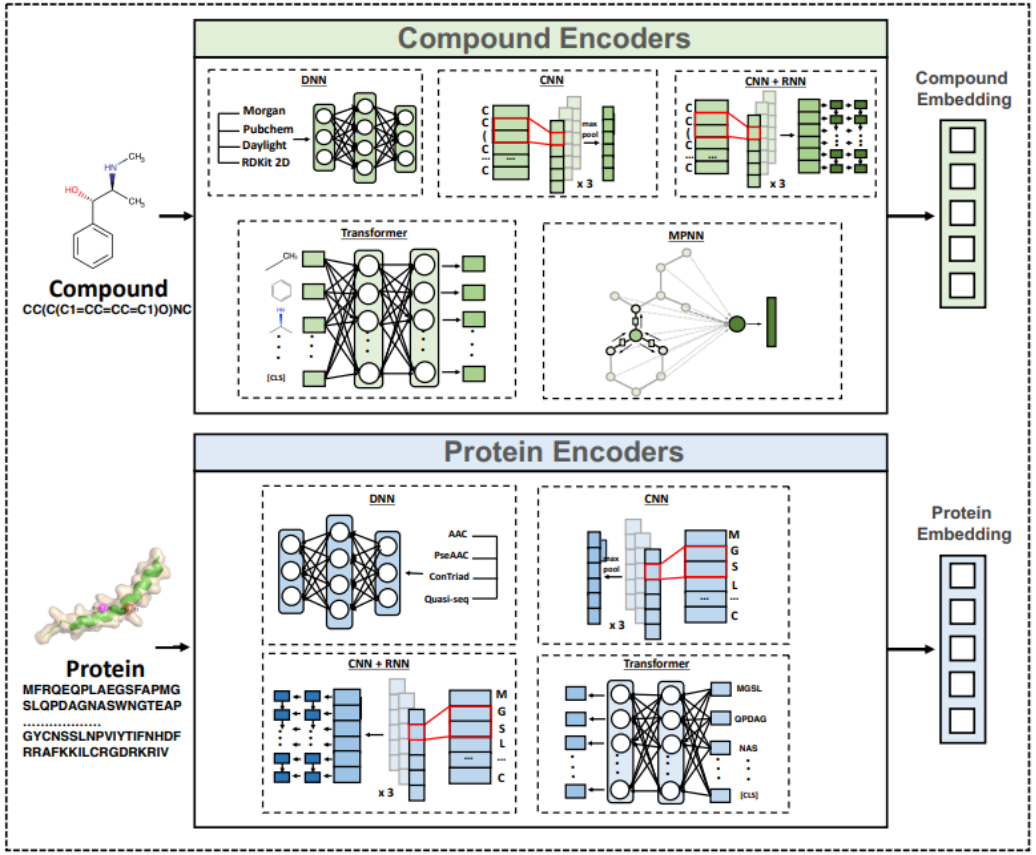
Prof. Zoran B. Djordjević

# Introduction

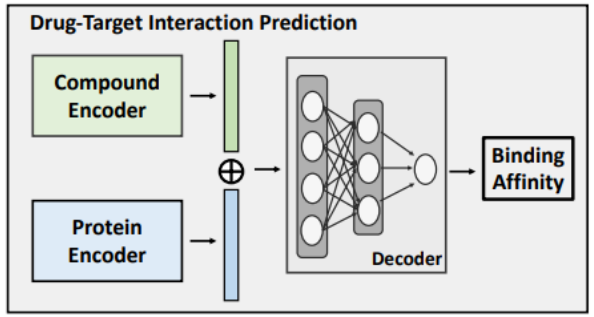
- DeepPurpose is a powerful deep learning library that specializes in drug discovery and drug repurposing tasks.
  - Drug-target interaction prediction
  - Compound property prediction
  - Protein-protein interaction prediction
  - Protein function prediction
  - Does not provide protein to protein amino acid sequence similarity
- The objective of the project is to leverage the built-in target encoders in DeepPurpose to encode protein sequence data from uniport and train a model to calculate the similarity pairwise scores.

# DeepPurpose

## A. Molecular Encoding Module




## B. Drug-Target Interaction Prediction




## C. DeepPurpose 10 Lines Framework

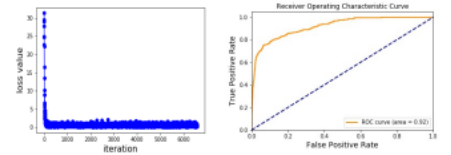
```
>>> from DeepPurpose import DTI
>>> from DeepPurpose.utils import *
>>> from DeepPurpose.dataset import *
>>>
>>> X_drug, X_target, y = load_process_DAVIS(SAVE_PATH, binary=False)
>>>
>>> drug_encoding, target_encoding = 'CNN', 'CNN'
>>> train, val, test = data_process(X_drug, X_target, y, drug_encoding, \
                                    target_encoding, split_method='random', \
                                    frac=[0.7,0.1,0.2], random_seed = 1)
>>>
>>> config = generate_config(drug_encoding, target_encoding, \
                             cls_hidden_dims = [1024,1024,512], \
                             train_epoch = 100, LR = 0.001, batch_size = 256, \
                             cnn_drug_filters = [32,64,96], \
                             cnn_drug_kernels = [4,8,12], \
                             cnn_target_filters = [32,64,96], \
                             cnn_target_kernels = [4,8,12])
>>>
>>> model = DTI.model_initialize(**config)
>>> model.train(train, val, test)
```

## D. SOTA Prediction Performance

| | Model | MSE | Concordance Index |
|---|---|---|---|
| **Dataset 1: DAVIS** | | | |
| Baselines | KronRLS | 0.329 (0.019) | 0.847 (0.006) |
| | GraphDTA | 0.263 (0.015) | 0.864 (0.007) |
| | DeepDTA | 0.262 (0.022) | 0.870 (0.003) |
| DeepPurpose | CNN+CNN | 0.254 (0.018) | 0.879 (0.011) |
| | MPNN+CNN | 0.271 (0.012) | 0.858 (0.007) |
| | MPNN+AAC | 0.242 (0.009) | 0.881 (0.005) |
| | CNN+Trans | 0.282 (0.009) | 0.852 (0.006) |
| | Morgan+CNN | 0.271 (0.012) | 0.858 (0.007) |
| | Morgan+AAC | 0.258 (0.012) | 0.861 (0.008) |
| | Daylight+AAC | 0.277 (0.014) | 0.861 (0.008) |
| **Dataset 2: KIBA** | | | |
| | Model | MSE | Concordance Index |
| Baselines | KronRLS | 0.852 (0.014) | 0.688 (0.003) |
| | GraphDTA | 0.183 (0.003) | 0.862 (0.005) |
| | DeepDTA | 0.196 (0.008) | 0.864 (0.002) |
| DeepPurpose | CNN+CNN | 0.196 (0.005) | 0.856 (0.004) |
| | MPNN+CNN | 0.222 (0.006) | 0.825 (0.003) |
| | MPNN+AAC | 0.178 (0.002) | 0.872 (0.001) |
| | CNN+Trans | 0.240 (0.013) | 0.818 (0.004) |
| | Morgan+CNN | 0.229 (0.008) | 0.825 (0.004) |
| | Morgan+AAC | 0.233 (0.009) | 0.823 (0.004) |
| | Daylight+AAC | 0.252 (0.014) | 0.808 (0.008) |

## E. More Functionalities

- **Training Monitoring and Metrics Auto-generation**




- **Repurposing and Screening Ranked List Generation**

Drug Repurposing Result for SARS-CoV2 3CL Protease

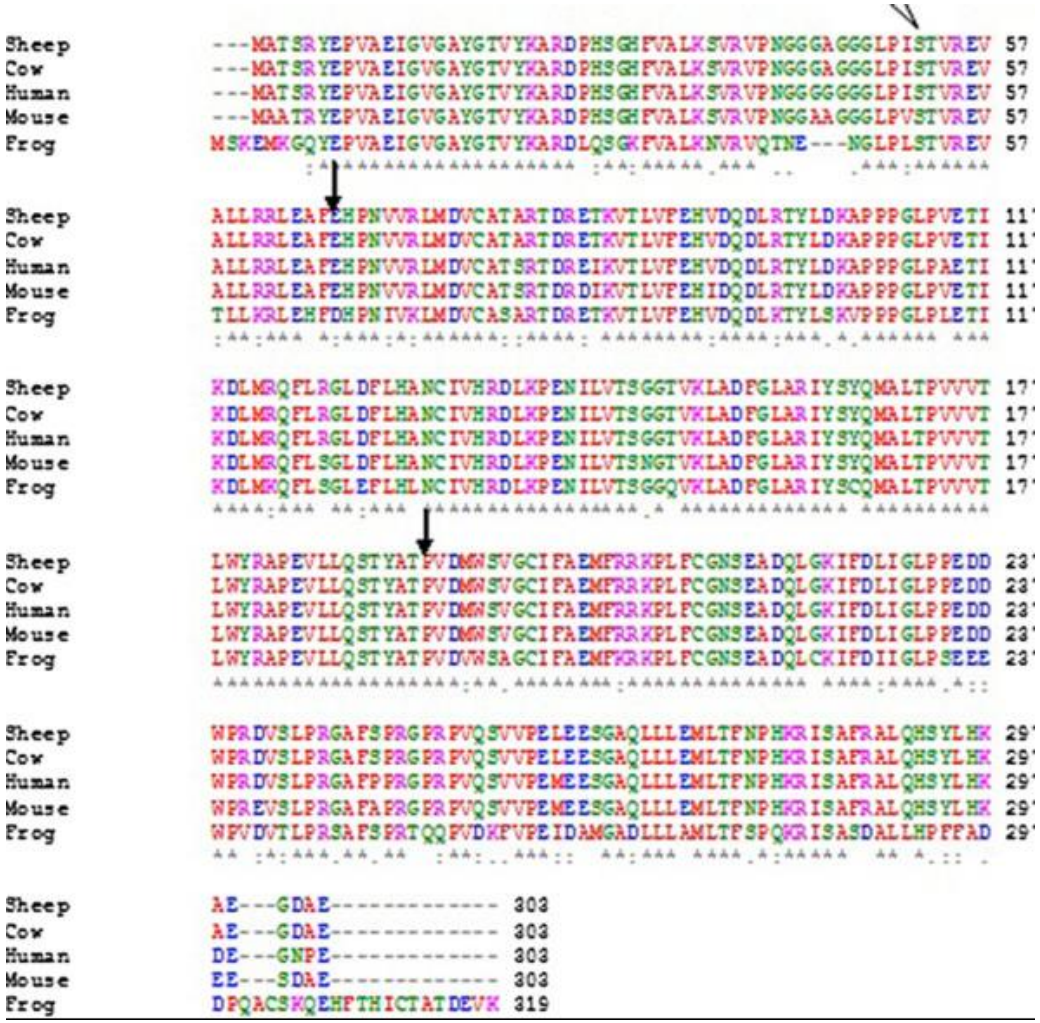
| Rank | Drug Name | Target Name | Binding Score |
|---|---|---|---|
| 1 | Sofosbuvir | SARS-CoV2 3CL Protease | 360.22 |
| 2 | Daclatasvir | SARS-CoV2 3CL Protease | 424.56 |
| 3 | Vicriviroc | SARS-CoV2 3CL Protease | 623.78 |
| 4 | Efavirenz | SARS-CoV2 3CL Protease | 768.33 |

- **Supporting Drug Property, Protein Function, Drug-Drug Interaction Prediction, and Protein-Protein Interaction Prediction, all less than 10 lines of codes.**

# DeepPurpose Target Encodings

| Target Encodings | Description |
|---|---|
| **AAC** | Amino acid composition up to 3-mers |
| **PseudoAAC** | Pseudo amino acid composition |
| **Conjoint_triad** | Conjoint triad features |
| **Quasi-seq** | Quasi-sequence order descriptor |
| **ESPF** | Explainable Substructure Partition Fingerprint |
| **CNN** | Convolutional Neural Network on target seq |
| **CNN_RNN** | A GRU/LSTM on top of a CNN on target seq |
| **Transformer** | Transformer Encoder on ESPF |

# Protein Sequence Similarity

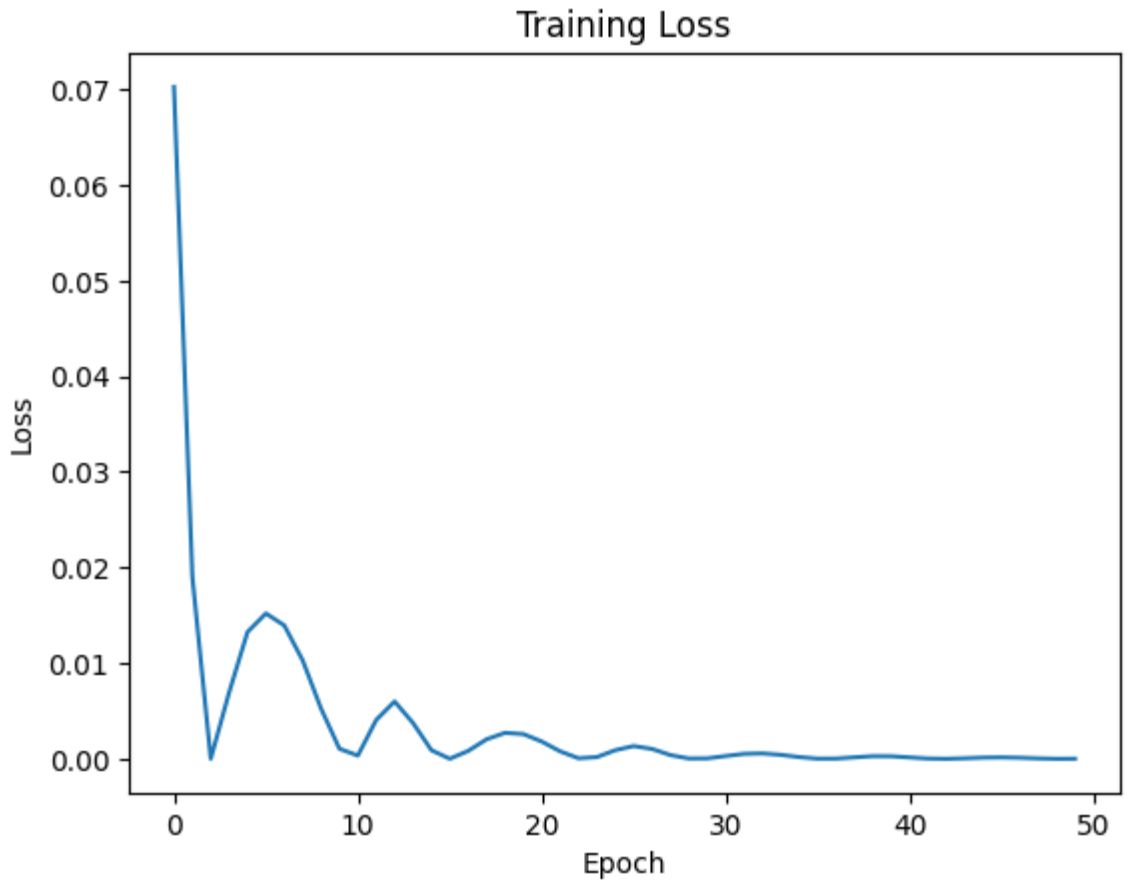


# Encoding Uniprot Protein Sequence Data

| | Target Sequence | target_encoding |
|---|---|---|
| **0** | MENERAKQVYLAKLNEQAERYDEMVEAMKKVAALDVELTIEERNLL... | [10.373, 5.809, 4.149, 5.809, 0.83, 12.033, 2.... |
| **1** | MENERAKQVYLAKLNEQAERYDEMVEAMKKVAALDVELTIEERNLL... | [9.804, 5.49, 4.706, 5.882, 0.784, 12.157, 3.1... |
| **2** | MSTREENVYMAKLAEQAERYEEMVEFMEKVAKTVDVEELSVEERNL... | [11.417, 4.724, 3.15, 5.906, 0.787, 12.598, 1.... |
| **3** | MAATLGRDQYVYMAKLAEQAERYEEMVQFMEQLVTGATPAEELTVE... | [10.623, 4.396, 2.93, 5.495, 1.099, 9.89, 3.66... |
| **4** | MATTLSRDQYVYMAKLAEQAERYEEMVQFMEQLVSGATPAGELTVE... | [10.976, 4.878, 2.439, 5.285, 0.813, 10.976, 3... |
| **...** | ... | ... |
| **94** | MDPEPTEHSTDGVSVPRQPPSAQTGLDVQVVSAAGDSGTMSQDTEV... | [7.434, 4.602, 3.894, 6.018, 0.354, 6.018, 5.1... |
| **95** | MIGARVFCITTTALRRSPIFFFPKIPTRPVFRLSPATRPIVAMSTT... | [6.542, 7.477, 3.738, 6.075, 2.336, 5.14, 4.20... |
| **96** | MQFLKSAKQKPNYYHIMLVEDQEEGTLHQFNYCERCSESQNNKCIS... | [5.426, 3.101, 4.91, 4.134, 3.876, 4.393, 5.16... |
| **97** | MFQAAVGPLQTNISLPEETPGLELNWAALLIVMVIIPTIGGNILVI... | [5.693, 4.95, 5.198, 2.475, 1.98, 4.455, 4.455... |
| **98** | MFQAAVGPLQTNISLPEETPGLELNWAALLIVMVIIPTIGGNILVI... | [8.475, 1.695, 2.825, 2.825, 1.695, 2.825, 4.5... |

# Similarity Model

```
# Instantiate the similarity scoring model
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(16840, 256),
    nn.ReLU(),
    nn.Linear(256, 64),
    nn.ReLU(),
    nn.Linear(64, 1),
    nn.Sigmoid()
)
```




```
# Calculate the similarity score for the pair of protein sequences
score = model(torch.cat((x1, x2), dim=1)).item()
print('Similarity score: {:.4f}'.format(score))
```

```
Similarity score: 0.7929
```

# Verification

**Our model predicted the similarity score of 0.7929**

**The NCBI Blast database score is 241/255 = 0.9450**

# Challenges

1. Data format challenge: DeepPurpose requires data in a specific format, but the documentation does not provide clear guidelines on this format. Users must either inspect their dataset or look at the code to determine the required format.

2. Large dataset challenge: The Uniprot protein sequence data is massive, with over 40,000 records, resulting in a possible 1.6 billion protein to protein pairs. Running a model on a good sample size requires a computer with high RAM, even when utilizing big data technologies such as Spark or Multiprocessing. The provided multiprocessing code in Python Notebook can help apply the similarity model on the full dataset, but most computers will run out of memory. This requires either a high-powered machine or chunking the data into batches and running the model on each batch.

3. Verification challenge: Our model similarity scores can be verified using NCBI Blast, either via a web interface or a Python library. However, using the Python library to verify large datasets of protein to protein pairwise similarity is very slow, as the API has a 15-second wait time after each protein-to-protein pair request. Each request response can take approximately 30 seconds. For example, sending an API request for 200 protein pairs can take approximately 1.5 to 2 hours.

# YouTube URLs

- 2 minutes video URL (https://youtu.be/791IqWVArco)

- 15 minutes video URL (https://youtu.be/Ie6s2vBdBfw)