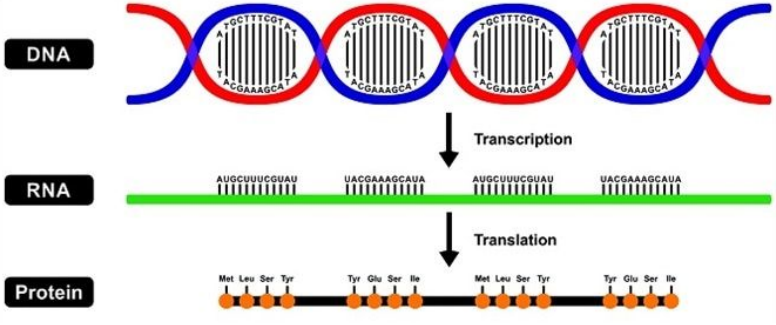
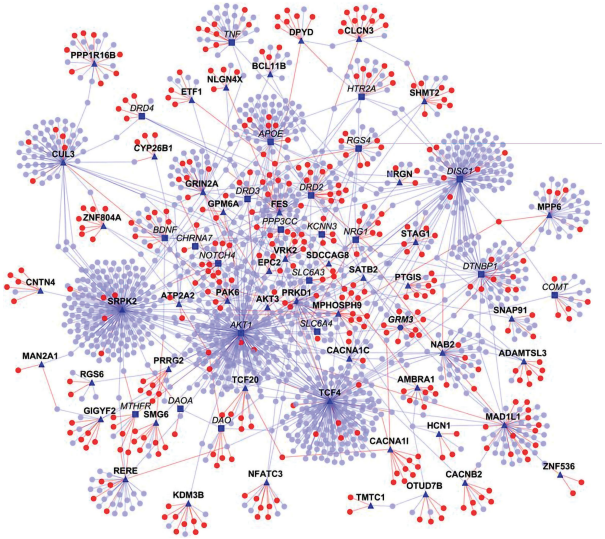
Deep Learning for Protein-Protein Interaction Prediction

Samuel Sledzieski

Protein Interaction Networks

- A lot of work has gone into genetic studies → how does the genome affects phenotype?
- It is equally understanding to gain a mechanistic understanding of biology → how do proteins interact?





Images Wikipedia/Shutterstock

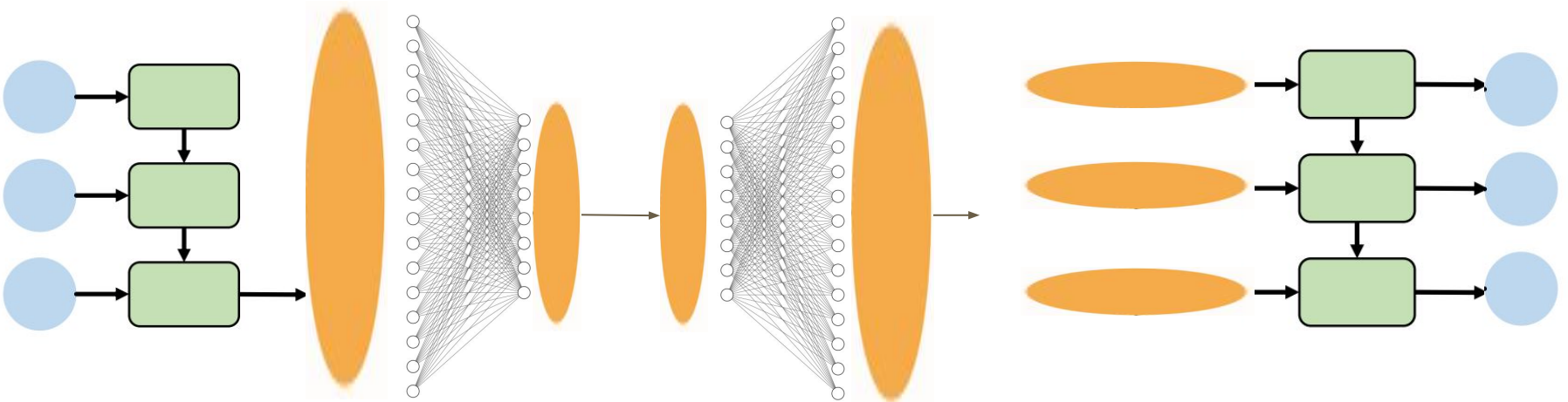
- Protein-Protein Interaction prediction from **sequence only** is a challenging task
- **Our contribution: we present a deep learning framework for PPI prediction**
  - Learn protein sequence embeddings with an **LSTM autoencoder**
  - Introduce the idea of an **interaction fingerprint**
  - Predict protein interaction using a **CNN**
  - Explore PPI **transfer learning** to other species

LSTM Autoencoder Architecture

Many to one LSTM

Dense layers
(256 → 128 → 256)

Many to many LSTM



Encoder

Decoder

Protein Interaction Embedding

Subsample data

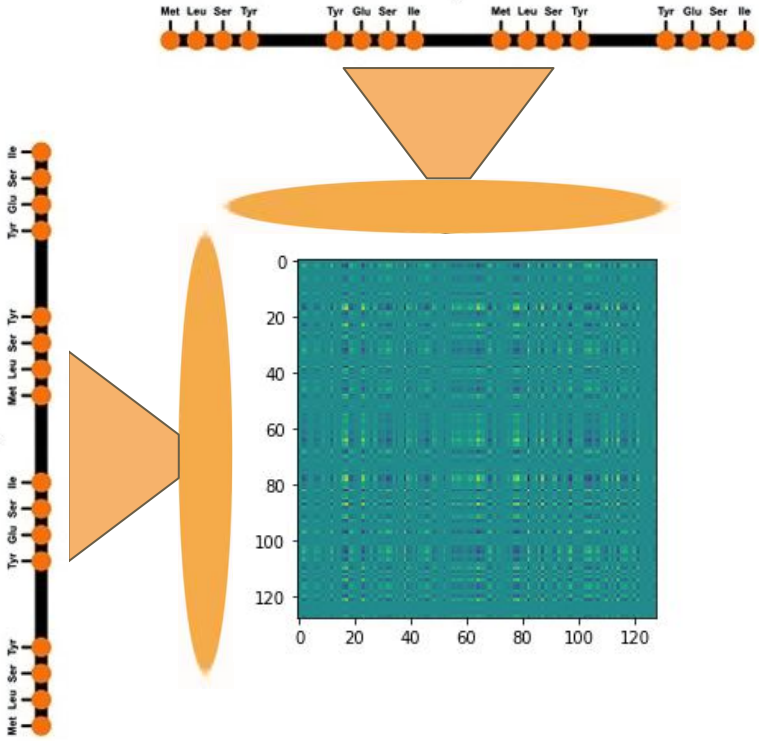
- *Homo sapiens* protein-protein interactions from STRING database
- Sample 2,000 true positive interactions
- Generate 13,000 negative interactions by randomly sampling pairs of proteins
    - Random interaction has ~3% chance of actually being true

Learn latent space of training data

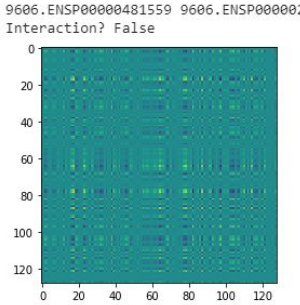
- Use trained LSTM-AE to generate a length 128 vector

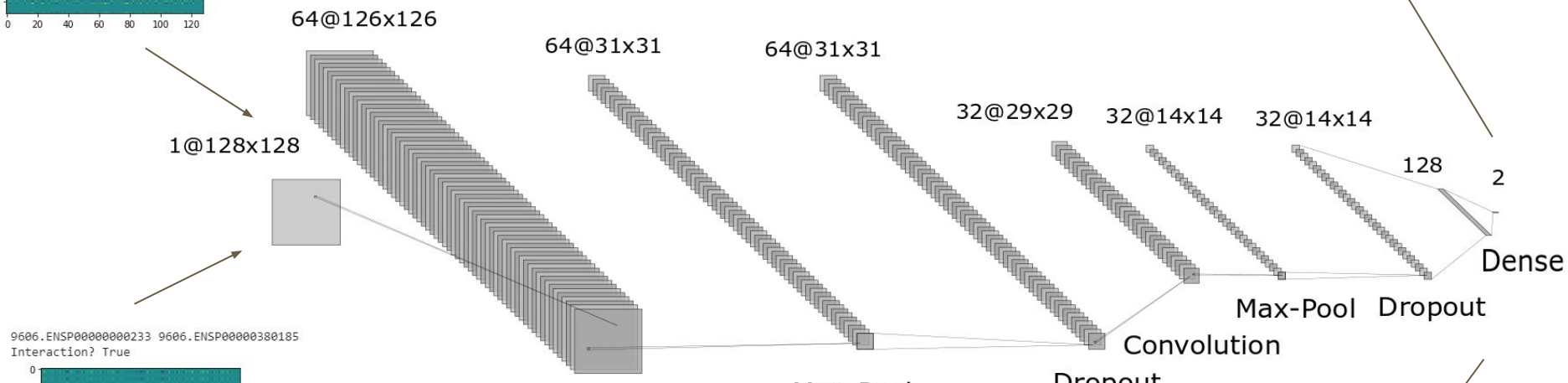
Generate interaction "fingerprints"

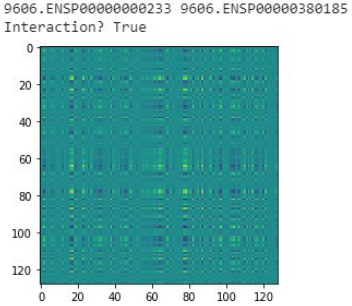
- Outer product of latent-space embeddings generates a unique image for each protein-protein interaction



CNN Architecture





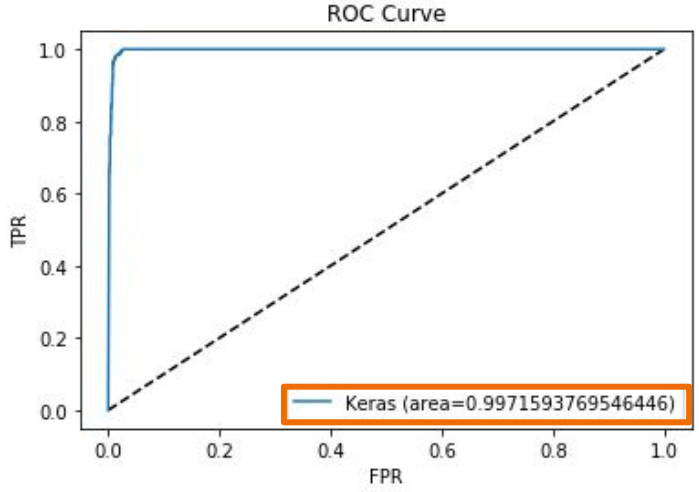


0	1
0.8991	0.1009

0	1
0.00001	0.99999

Model Training & Evaluation

- 75-25 Train-Test Split
- Trained LSTM-AE on 1,000 samples for 5 epochs, batch size 32
- Trained CNN on 11,251 samples (2,000 positive, 13,000 negative) for 50 epochs, batch size 128
- Evaluated CNN on 3,750 never-seen samples
- Achieved 96.24% training accuracy, 94.13% testing accuracy, 0.997 AUROC



Takeaways:

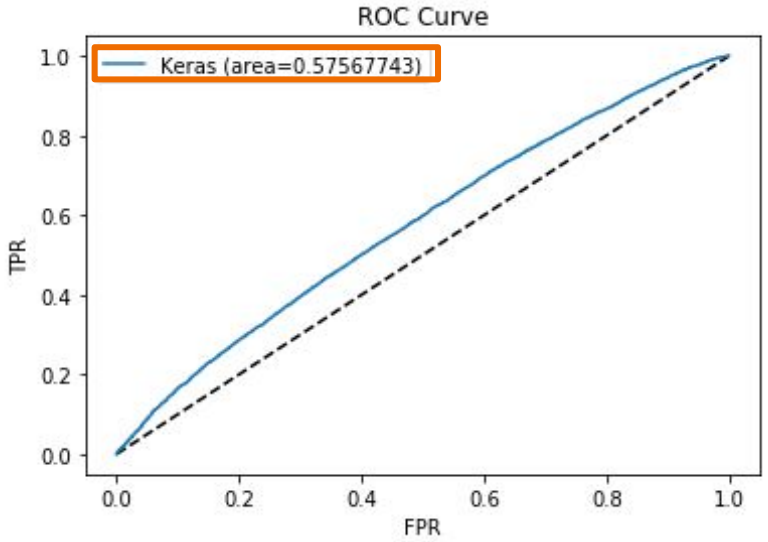
- LSTM-AE training is **slow**
- CNN training is **fast**
- Embedding + outer product aids prediction of protein interaction

```
Epoch 45/50
11251/11251 [==============================] - 2s 204us/sample - loss: 0.1224 - accuracy: 0.9525
Epoch 46/50
11251/11251 [==============================] - 2s 203us/sample - loss: 0.1196 - accuracy: 0.9548
Epoch 47/50
11251/11251 [==============================] - 2s 204us/sample - loss: 0.1135 - accuracy: 0.9556
Epoch 48/50
11251/11251 [==============================] - 2s 202us/sample - loss: 0.1152 - accuracy: 0.9539
Epoch 49/50
11251/11251 [==============================] - 2s 201us/sample - loss: 0.1095 - accuracy: 0.9571
Epoch 50/50
11251/11251 [==============================] - 2s 202us/sample - loss: 0.1022 - accuracy: 0.9624

3750/3750 [==============================] - 0s 117us/sample - loss: 0.1440 - accuracy: 0.9413
Test Loss, Test Accuracy: [0.14399235029617946, 0.94133335]
```

Generalization to other species

- Can we apply our trained models to data from different species?
  - *Saccharmoyces Cerevisiae* (Yeast)
  - *Drosophilia melanogaster* (Fruit fly)
- Sampled 10,000 positive interactions between the two species
- Sampled 10,000 negative interactions with one protein coming from each species
- Embed proteins and predict interaction using pre-trained model
- Use same embeddings but retrain CNN



Takeaways:

- Accuracy is **poor** using same embedding and CNN
- Accuracy **improves** using the same embeddings but retraining the CNN
- Embeddings, and to an extent prediction models, are difficult to transfer across species
- There may be potential for weight sharing / inductive transfer learning to speed training

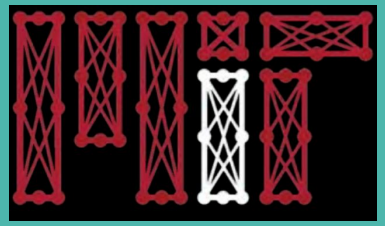
Download the models and data

shorturl.at/buyF1

Download the models only

shorturl.at/acCFL

https://github.com/samsledje/Deep_PPI



References in Github README