Predicting Protein-Protein Interactions

1/20/21

Nikhil Lonberg

Overview

Introduction

Problem Statement → The Interactome → Assumptions → Constraints

Methods

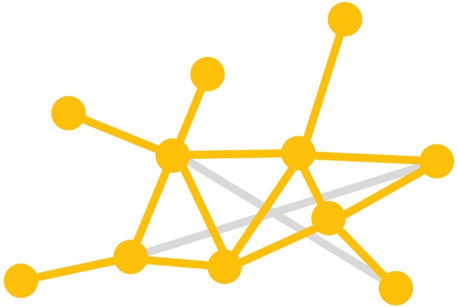
Datasets → Model Design → Encoders → Hyperparameters

Results

Conjoint Triad → Autocovariance → Res2Vec

Discussion

Findings → Applications → Future Directions



Introduction

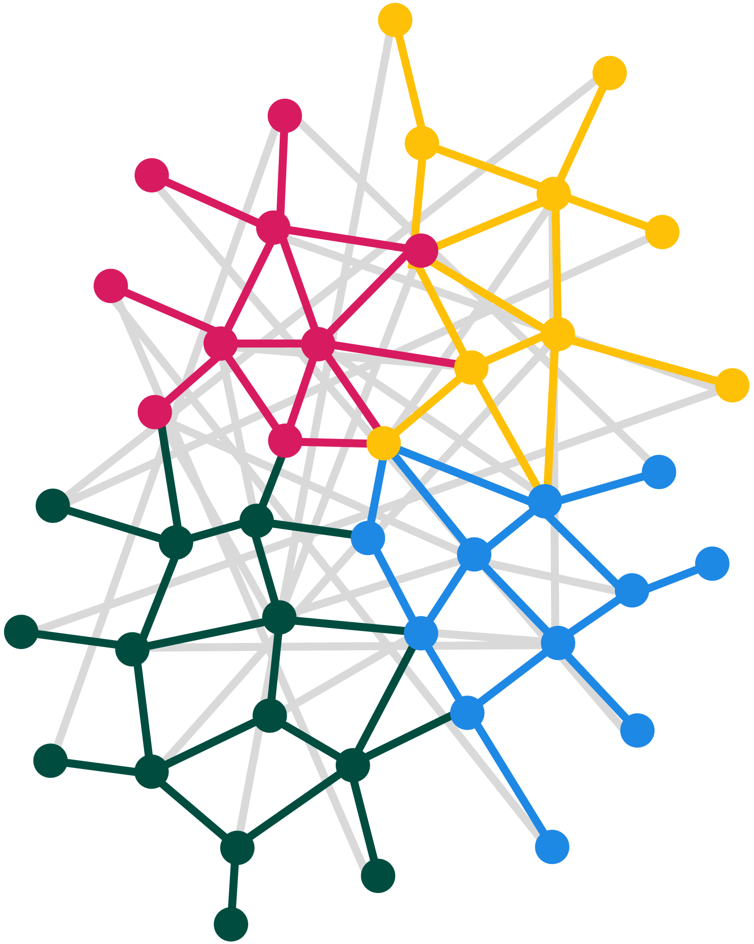
Problem Statement → The Interactome → Assumptions → Constraints

Problem Statement

Given a pair of proteins' amino acid sequences, predict whether those two proteins will interact.

The Interactome

- Researchers collaborating to reconstruct protein interaction network that underlies all cellular function.
- There are many definitions of protein “interactions”.
- Computer-simulated PPI prediction supplements wet lab methods; reduces costs and time.



Assumptions

1.) Amino acid sequences alone provide enough information to accurately model PPIs.

2.) Proteins either interact or do not interact.

3.) Deep Learning is essential to uncovering complex relationships in protein data.

Assumptions

1.) Amino acid sequences alone provide enough information to accurately model PPIs.

Risks: PPIs are affected by temperature, pressure, dissolved ions, and post-translational modifications not accounted for by sequence data.

Rationale: AA sequences are widely available and almost entirely determine structure and function of proteins.

Assumptions

2.) Proteins either interact or do not interact.

Risks: UniProt’s PPI data is a binary classification. Light affinities between proteins are likely classified as “Not Interacting”. But we do not know the experimental methods that went into these classifications, so we are assuming a lot of error that we do not know anything about.

Rationale: Binding affinities are scarcely available. Also, deep learning is not typically used for regression.

Assumptions

3.) Deep Learning is essential to uncovering complex relationships in protein data.

Risks: Neural networks reduce our ability to make inferences and dramatically increase training time.

Rationale: Research indicates that multilayer neural networks are quite good at finding hidden spatial relationships.

Constraints

Tier One: Accuracy

Tier Two: Simplicity, Speed

Tier Three: Inference

Constraints

Tier One: Accuracy

Deep Learning ↑

Tier Two: Simplicity, Speed

Deep Learning ↓

Tier Three: Inference

Deep Learning ↓

Constraints

Tier One: Accuracy

Deep Learning ↑ AA Sequences ↓

Tier Two: Simplicity, Speed

Deep Learning ↓ AA Sequences ↑

Tier Three: Inference

Deep Learning ↓ AA Sequences ↑

Constraints

Tier One: Accuracy

Deep Learning ↑ AA Sequences ↓ Binary Interaction ↓

Tier Two: Simplicity, Speed

Deep Learning ↓ AA Sequences ↑ Binary Interaction ↑

Tier Three: Inference

Deep Learning ↓ AA Sequences ↑ Binary Interaction ↓

Constraints

Tier One: Accuracy

Deep Learning ↑ AA Sequences ↓ Binary Interaction ↓

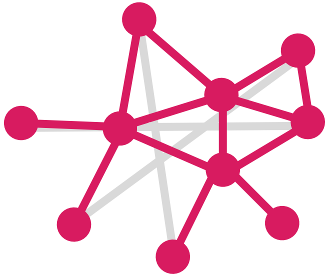
Tier Two: Simplicity, Speed

Deep Learning ↓ AA Sequences ↑ Binary Interaction ↑

Tier Three: Inference

Deep Learning ↓ AA Sequences ↑ Binary Interaction ↓

Conclusion: Constraints in tension; assumptions mean compromises.



Methods

Datasets → Model Design → Encoders → Hyperparameters

Datasets

Training dataset

- 52,310 pairs of human proteins.
- 57% of protein pairs in training set interact and 42% of pairs do not interact.

Testing datasets

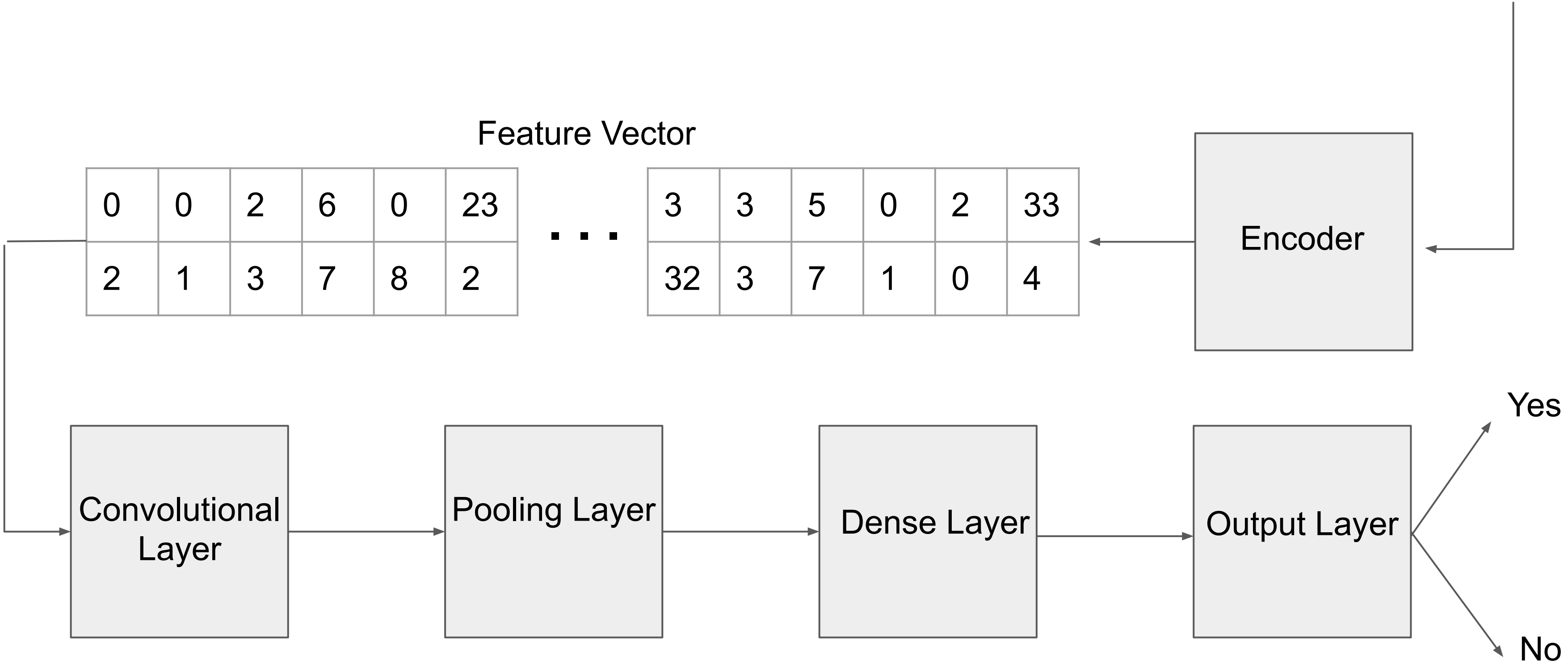
- 900 pairs of SARS-CoV-2 and human proteins; 1819 pairs of yeast proteins.
- 52% of protein pairs in testing set interact and 42% of pairs do not interact.

Transfer learning

- Res2Vec feature embeddings trained on 556,127 protein sequences

Model Design

Protein 1 Sequence: AGKEDPS...WYIKCCOI
Protein 2 Sequence: VTSW1RK...DEYMPHF



Feature Vector

0	0	2	6	0	23	. . .	3	3	5	0	2	33
2	1	3	7	8	2		32	3	7	1	0	4

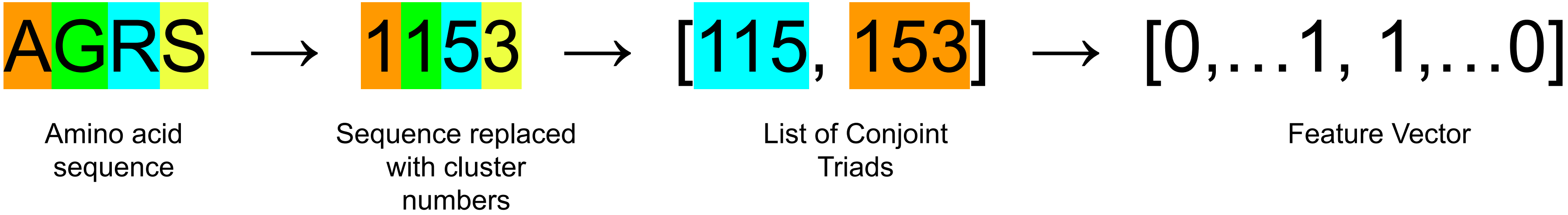
Conjoint Triad Encoder

From “Predicting protein-protein interactions based only on sequences information.” Shen et. al. 2007.

Step One: cluster amino acids according to dipole and side chain volume.

Step Two: replace amino acid labels with cluster numbers.

Step Three: count triads with sliding scale.



AGRS → 1153 → [115, 153] → [0,…1, 1,…0]

Amino acid sequence

Sequence replaced with cluster numbers

List of Conjoint Triads

Feature Vector

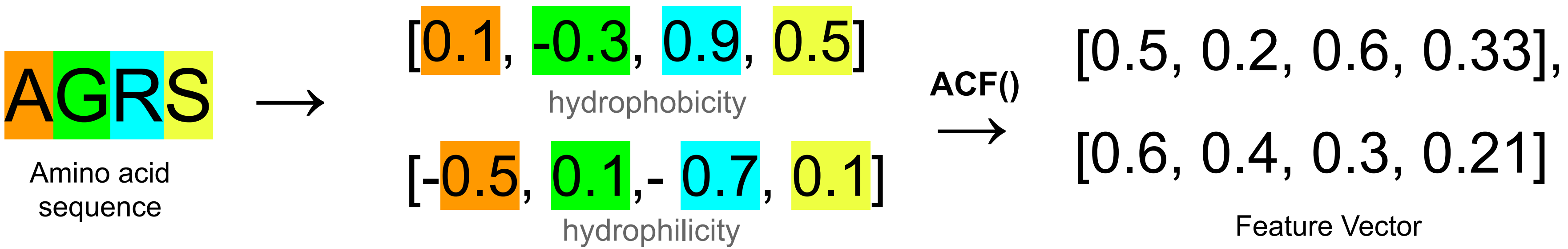
Autocovariance Encoder

From “Sequence-based prediction of protein protein interaction using a deep-learning algorithm” Sun et. al. 2017.

Step One: collect data on amino acid properties.

Step Two: replace amino acid labels with property values.

Step Three: perform autocovariance analysis sequences of values.



AGRS

Amino acid sequence

→

[0.1, -0.3, 0.9, 0.5]

hydrophobicity

[-0.5, 0.1,- 0.7, 0.1]

hydrophilicity

ACF()

→

[0.5, 0.2, 0.6, 0.33],

[0.6, 0.4, 0.3, 0.21]

Feature Vector

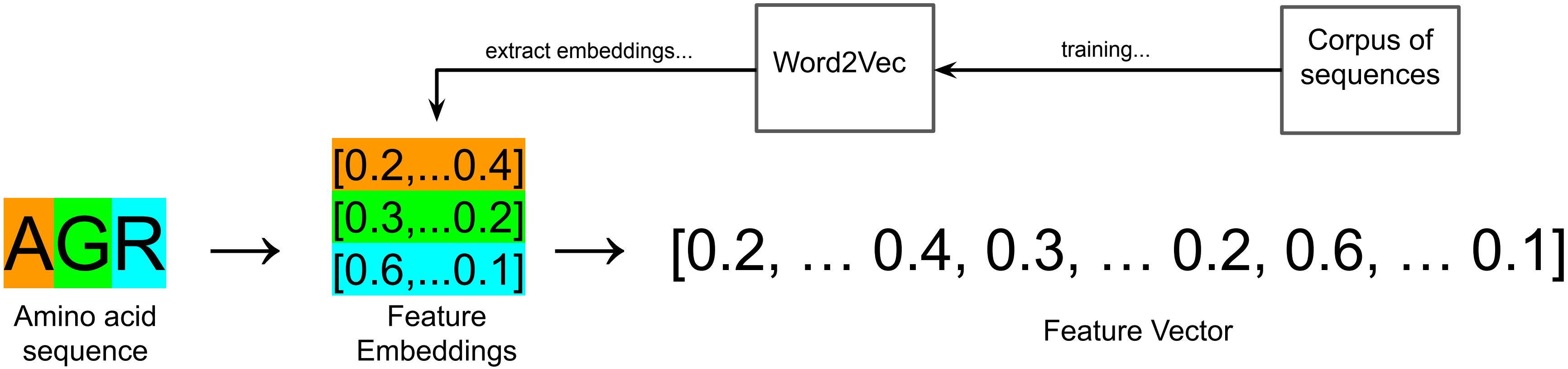
Res2Vec Encoder

From “Integration of deep learning with feature embedding for protein–protein interaction prediction.” Yao et.al. 2019.

Step One: train Word2Vec NLP model on corpus of sequences.

Step Two: extract feature embeddings from model.

Step Three: replace AA labels with feature embeddings.



extract embeddings...

Word2Vec

training...

Corpus of sequences

AGR →

Amino acid sequence

[0.2,...0.4]
[0.3,...0.2]
[0.6,...0.1]

→

Feature Embeddings

[0.2, … 0.4, 0.3, … 0.2, 0.6, … 0.1]

Feature Vector

Hyperparameters

Encoders: Window size, number of clusters, cluster properties, autocovariance lag, “word” size, embeddings vector length

CNN: convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

2

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

$$2^{10} =$$

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

$$2^{10} = 1024 \text{ models}$$

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

2^{10} = 1024 models

1024 * 5 =

Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

2^{10} = 1024 models

1024 * 5 = **5120 runs**

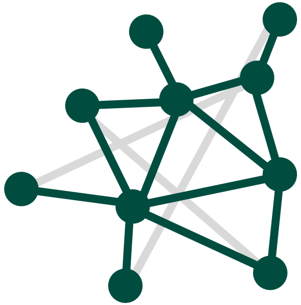
Hyperparameters

Example GridSearch Hyperparameters: window size, number of clusters, cluster properties, convolutional layers, hidden layers, epochs, batch size, dropout, nodes per layer, kernel size

2^{10} = 1024 models

1024 * 5 = **5120 runs**

Conclusion: Gridsearch is too performance exhaustive.

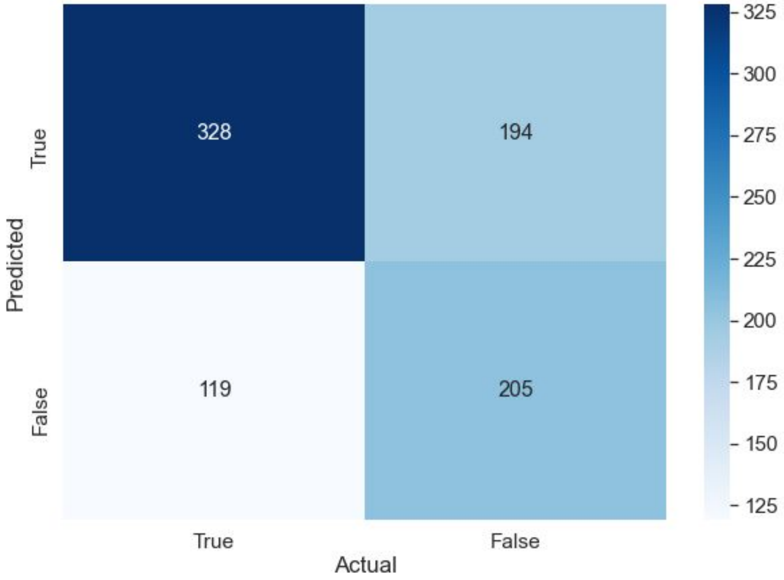


Results

Conjoint Triad Method → Autocovariance → Res2Vec

Conjoint Triad Model Evaluation

SARS-CoV-2 Confusion Matrix



Baseline Accuracy: 57%

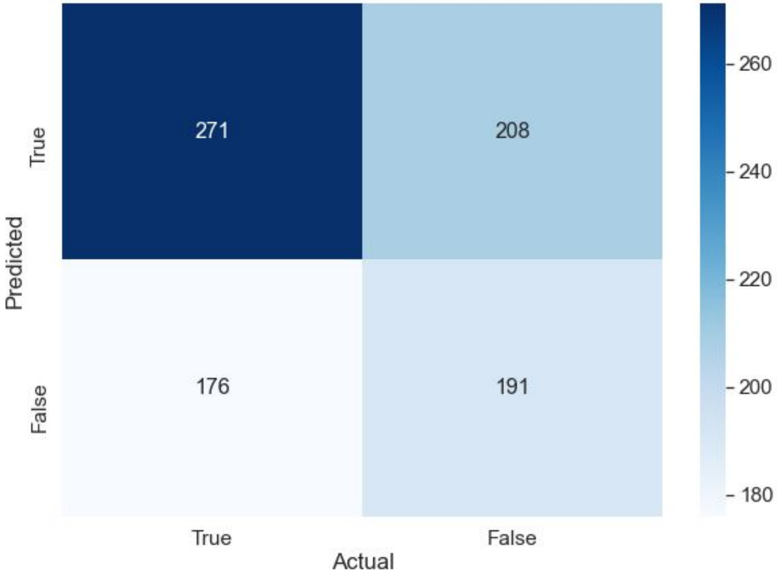
Model Accuracy: 77.5%*

	Accuracy	**Specificity**	**Sensitivity**	**Precision**
Human	77.5% *	n/a	n/a	n/a
Yeast	54.2%	68.4%	39.1%	53.9%
SARS-Cov-2	63.0%	51.4%	73.4%	62.8%

* Average accuracy across 5-fold cross-validation of training set.

Autocovariance Model Evaluation

SARS-CoV-2 Confusion Matrix



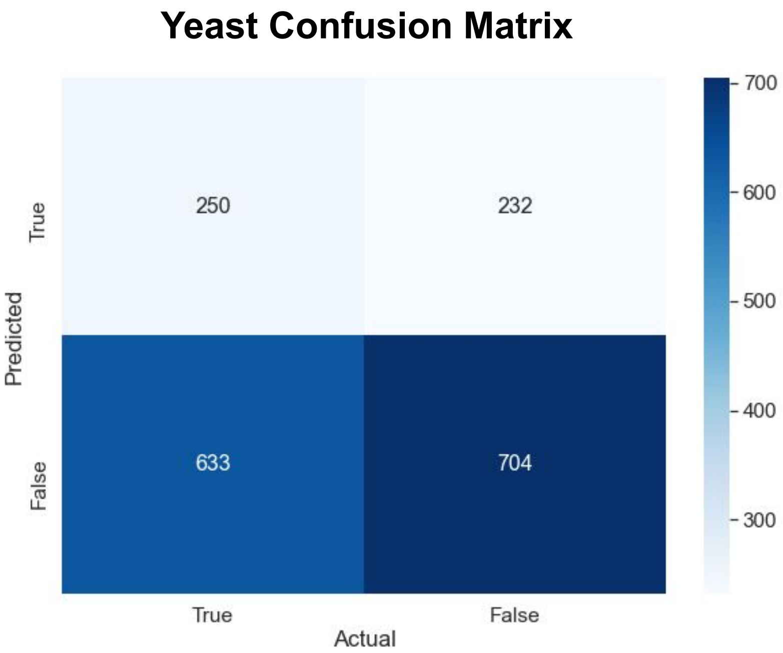
Baseline Accuracy: 57%

Model Accuracy: 66.5%*

	Accuracy	Specificity	Sensitivity	Precision
Human	66.5% *	n/a	n/a	n/a
Yeast	52.4%	68.4%	35.9%	51.5%
SARS-Cov-2	54.6%	47.9%	60.6%	56.6%

* Average accuracy across 5-fold cross-validation of training set.

Res2Vec Model Evaluation

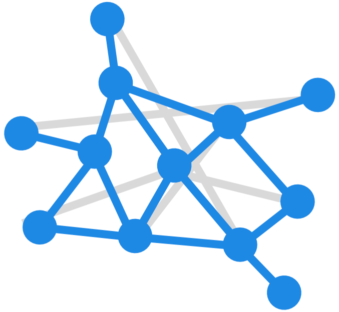


Baseline Accuracy: 57%

Model Accuracy: 79.0%*

	Accuracy	Specificity	Sensitivity	Precision
Human	79.0% *	n/a	n/a	n/a
Yeast	52.4%	75.2%	28.3%	51.9%
SARS-Cov-2	66.3%	52.1%	75.6%	60.3%

* Average accuracy across 5-fold cross-validation of training set.



Discussion

Findings → Applications → Future Directions

Findings

- Res2Vec performs best among encoders: 22% increase in accuracy over baseline.
- No models generalize to non-human protein datasets
- Error from heterogeneous datasets
- Hyperparameter search is too performance-intensive
- Success of model depends on task

Applications

- Discovering function of orphan proteins.
- Supplementing experimental methods.
    - Example: Weeding out false positives of co-immunoprecipitation assay.
- Building out protein-protein interaction networks.

Future Directions

- More practical hyperparameter tuning methods: Random Search, Bayesian Optimization, etc.
- Recurrent Neural Networks (e.g. Long Short Term Memory)
- Train on larger datasets and PPI data from other databases
- Focus model on more specific problem