Out-of-distribution multi-organism protein-protein interaction prediction for peptide therapeutics

Joseph **Szymborski**[1,2] and Amin **Emad**[1,2,3]

1. Department of Electrical and Computer Engineering, McGill University, Montréal, QC Canada
2. Mila, Québec AI Institute, Montréal, QC Canada
3. The Rosalind and Morris Goodman Cancer Institute, Montréal, QC Canada

1. Introduction

1.1 Motivation

- Protein-Protein Interaction (PPI) networks for model organisms are getting **ever-larger**.
- The size of PPI networks of species which are **not model organisms** is **much smaller**.

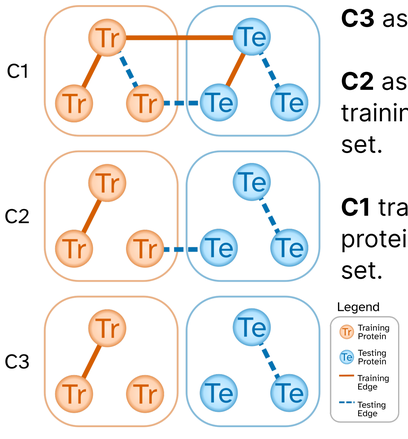
1.2 Out-of-Distribution Predictions

- The organisms with the **most incomplete graphs** often have **too little data** to train good PPI prediction models.
- It's desirable to therefore **train on model organisms** and **testing on species with little data**.
- Many ML models **do not make accurate out-of-distrubution** (OOD) predictions [1].

2. Methodology

2.2 Special Considerations for Validation & Testing Dataset

Park & Marcotte identified an information leakage problem with PPI prediction validation techniques [1].

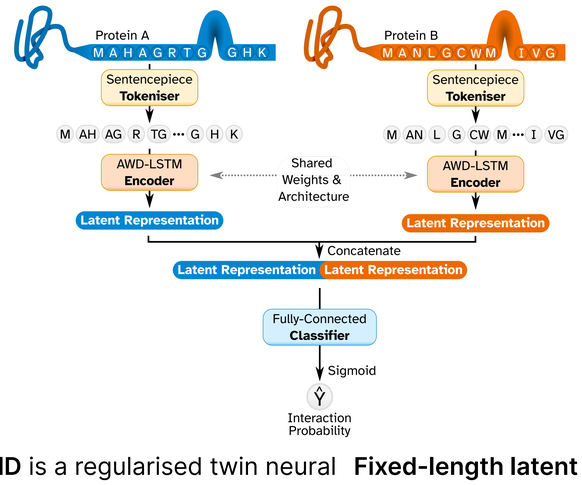


C3 assures no proteins in the testing set.

C2 assure no more than one protein in training pairs are in the testing or validation set.

C1 training pairs may contain one or two proteins found in the testing or validation set.

2.3 Overview of the RAPPPID Architecture



RAPPPID is a regularised twin neural network that adopts a modified AWD-LSTM [2].

RAPPPID considers **pairs of amino acid (AA) sequences** with an interaction label [3].

AA sequences are first tokenised with the **Sentencepiece algorithm** [4].

Fixed-length latent vector representations are computed for each sequence using **bi-directional AWD-LSTMs**.

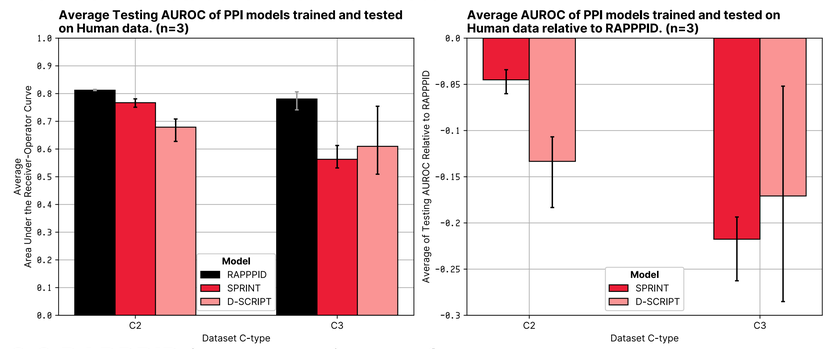
Latent vectors are **concatenated and are inputted** into a two-layer fully-connected **classification head**.

Output of the classifier is the **interaction probability**.

3. Results

3.1 RAPPPID generalises better than leading PPI prediction methods.

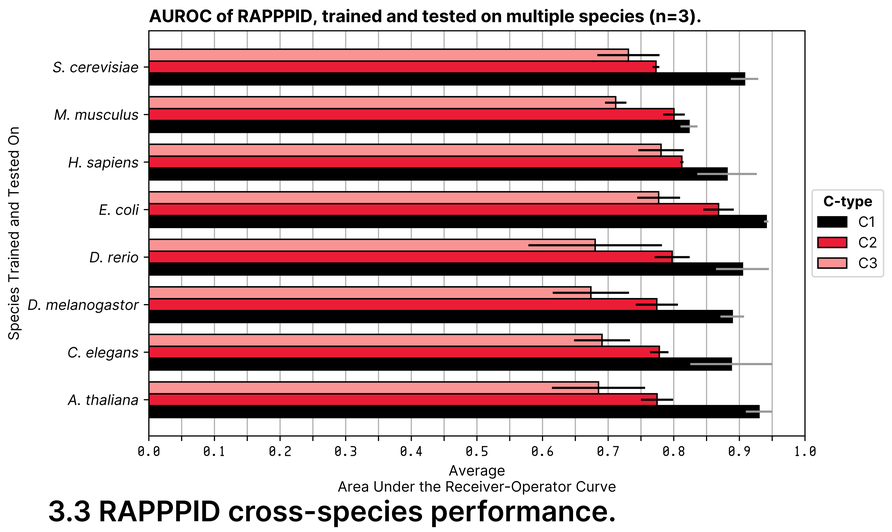
Across C2 and C3 *Homo sapiens* datasets, **RAPPPID achieves a higher AUROC than all other methods tested**.



3.2 RAPPPID intra-species performance.

RAPPPID models trained and tested on various species maintain a high degree of performance.

The average AUROC of three different random seeds and data splits are reported.



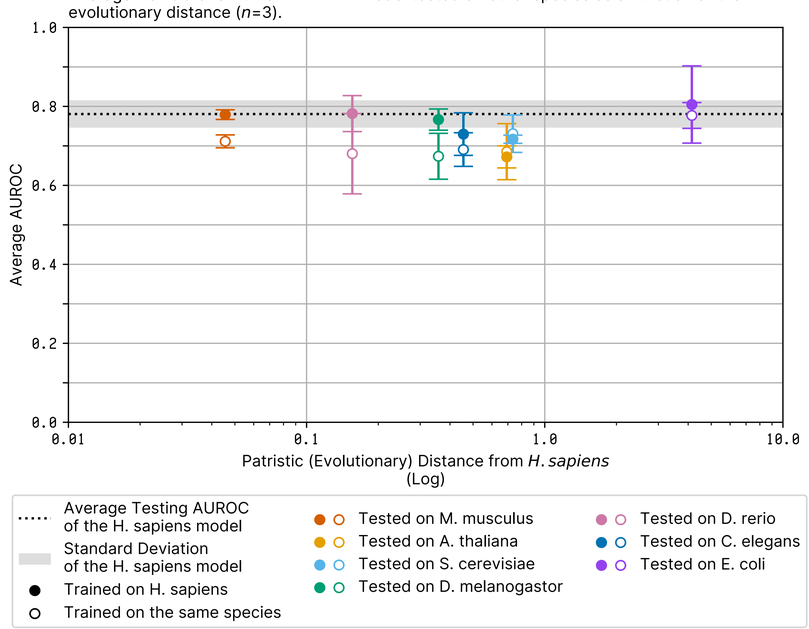
3.3 RAPPPID cross-species performance.

The performance of a C3 RAPPPID model trained on *H. sapiens* PPI data was measured on testing sets composed of PPIs of other species.

This was compared to RAPPPID trained & tested on the same species.

RAPPPID trained on Human PPIs maintains its performance when tested on PPIs of other species.

Average AUROC of a Human RAPPPID model tested on other species as a function of their evolutionary distance (*n*=3).



3. Results

3.4 RAPPPID transfer-learning performance.

To evaluate how RAPPPID out-of-distribution performance changes with fine-tuning, RAPPPID was trained on *H. sapiens*, and fined-tuned on *Escherichia coli* PPI data. Finally, the model was tested on *E. coli* PPIs.

The means of three different random seeds and data slits are reported below.

Training Species	Testing Species	Fine-Tuning Species	AUROC (Mean)	APR (Mean)
E. coli	E. coli	None	0.818	0.840
H. sapiens	E. coli	None	0.839	0.877
H. sapiens	E. coli	E. coli	**0.867**	**0.890**

4. Future Directions

4.1 Online full-proteome prediction server.

We are currently working on developping an online server for PPI prediction using RAPPPID.

We will leverage RAPPPID's efficiency to enable the predition of an inputted amino acid sequence against entire known proteomes of a specified organism.

4.2 Tools for therapeutic peptide discovery.

RAPPPID's out-of-distribution performance on unseen proteins across species lends itself quite naturally to the context of therapeutic peptide discovery.

We're developping tools to aid in the discovery of therapeutic peptides using RAPPPID's online interface.

5. Acknowledgments

Mila
Digital Research Alliance of Canada
Alliance de recherche numérique du Canada
INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION
FONDATION CANADIENNE POUR L'INNOVATION
Calcul Québec
VADASZ SCHOLARS

6. References

1. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods. 2012 Dec;9(12):1134–6.

2. Merity S, Keskar NS, Socher R. Regularizing and Optimizing LSTM Language Models. arXiv:170802182 [cs] [Internet].

3. **Szymborski** J, **Emad** A. RAPPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. Bioinformatics. 2022 Aug 15;38(16):3958–67.

4. Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: EMNLP 2018

See more information online

https://jszym.com/meetings/2023_cshl