RAPPPID: Towards Generalisable Protein Interaction Prediction with AWD-LSTM Twin Networks

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

[1]Department of Electrical and Computer Engineering, McGill University, Montréal, QC, Canada
[2]Mila, Quebec AI Institute, Montréal, QC, Canada
[3]The Rosalind and Morris Goodman Cancer Institute, Montréal, QC, Canada
*amin.emad@mcgill.ca

1. Introduction

1.1 Motivation

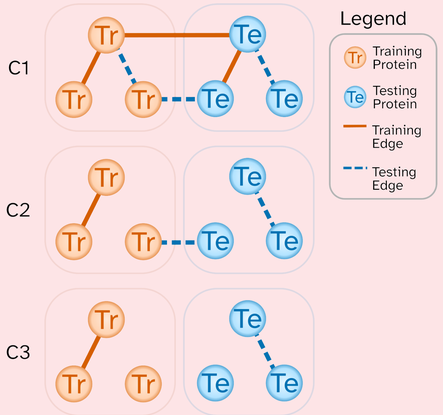
- Uncovering protein-protein interactions (PPIs) is very important for understanding most biological processes.
- Interactions can be **validated by a number of experiments**, however **they are costly** in terms of time, labour, and materials [1].
- **Computational approaches** to predict protein-protein interactions (PPIs) are therefore useful to help towards **reducing the number of costly experiments** researchers are required to perform.

1.2 Information Leakage in PPI Datasets

- The nature of PPI networks makes it easy to create datasets with **testing/training splits which leak information** [2].
- This results in **inflated performance metrics** that cannot properly assess the generalisability of these methods.

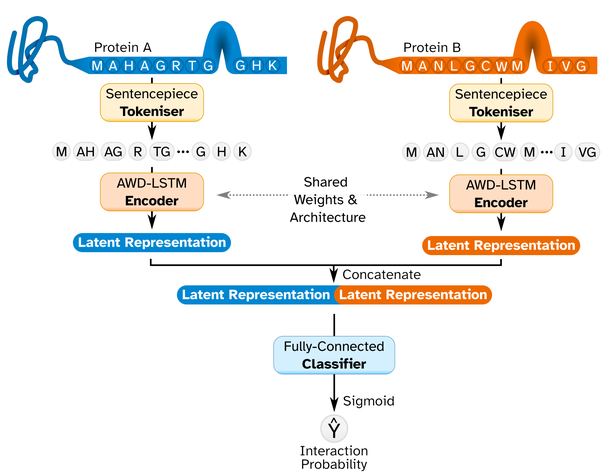
2. Methodology

2.1 Special Considerations for Validation & Testing Dataset Construction



- Park & Marcotte identified an information leakage problem with PPI prediction validation techniques [2].
- They described three types of validation sets (C1, C2, and C3).
- **C3** assures no proteins in the testing or validation set are in the training set.
- **C2** assures no more than one protein in training interaction pairs are in the testing or validation set.
- **C1** training pairs may contain one or two proteins found in the testing or validation set.

2.2 Overview of the RAPPPID Architecture

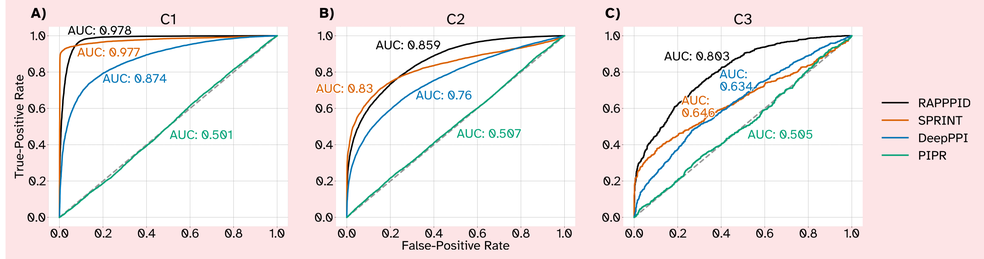


- RAPPPID is a regularised twin neural network that adopts a modified AWD-LSTM [3].
- RAPPPID considers pairs of amino acid (AA) sequences with an interaction label.
- AA sequences are first tokenised with the Sentencepiece algorithm [4].
- Fixed-length latent vector representations are computed for 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

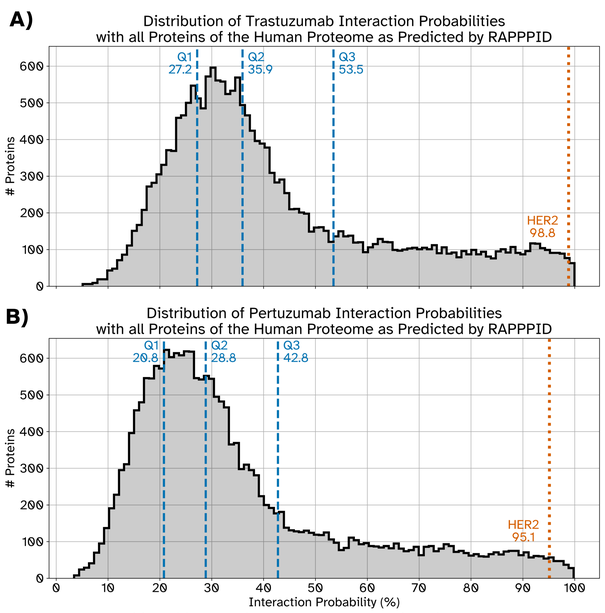
3. Results

3.1 Performance evaluation of RAPPPID and other algorithms

- Across C1, C2, and C3 testing datasets, **RAPPPID achieved higher AUROC than all other methods tested**.
- The margin between RAPPPID and the second highest performing method (SPRINT in all cases) was **highest when performed on the stricter C3 dataset**, resulting in approximately a **24.3% improvement**.
- The improvement obtained by RAPPPID compared to SPRINT was **lower on the C2 dataset (approximately 3.4%)**, and finally **nearly equivalent on the least strict C1 dataset**.
- Experiments were also conducted to establish the independence of RAPPPIDs accuracy and the similarity between the sequences evaluated.
- To further isolate any effects on model performance from the dataset, **we repeated the experiment on multiple random training, testing, and validation splits** as well as stratifying model performance by PPI evidence.
- All experiments indicated that model performance was not unduly influenced by our treatment of the dataset.



3.2 RAPPPID predicts interaction of HER2 with Trastuzumab and Pertuzumab



The probability of interaction between the antibodies (Trastuzumab and Pertuzumab) and each protein in the human proteome (according to STRING) were individually predicted using RAPPPID.

- Peptides and proteins have emerged as an important class of therapeutics, enabling researchers to target previously "undruggable" targets.
- We wanted to show how RAPPPID can be used to validate the interaction between candidate peptides and their targets.
- Here, we use RAPPPID to test existing therapeutics (Trastuzumab and Pertuzumab) against their known targets (HER2).
- Trastuzumab and Pertuzumab are recombinant humanised monoclonal antibodies used for the treatment of certain metastatic breast cancers.
- RAPPPID predicts the interaction probability between HER2 and Trastuzumab to be 98.8% and 95.1 with Pertuzumab.
- When compared to the probability of all other human proteins, the interaction between HER2 and Trastuzumab is in the first percentile, and in the second for Pertuzumab.

4. Conclusion

- RAPPPID succesfully addresses the challenges of creating generalisable PPI prediction models posed by inherent characteristics of PPI datasets.
- By adopting a modified AWD-LSTM training routine, RAPPPID was able to surpass state-of-the-art models under testing conditions that carefully controlled for information leakage and other sources of prediction accuracy inflation.
- RAPPPID's ability to predict interactions warrants further study into relevant tasks that might benefit from a similar approach.

5. Acknowledgments

Mila
compute canada | calcul canada
Calcul Québec
MiCM McGill initiative in Computational Medicine
INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
VADASZ SCHOLARS

6. References

More information including references can be found at **https://jszym.com/meetings/2022_ismb**