RAPPPID: Improving Protein Interaction Prediction on Unseen Proteins

SCAN ME

Joseph Szymborski & Amin Emad

Intelligent Systems for Molecular Biology
2022 Madison, Wisconsin, USA

McGill

Emad's Computational Biology and Artificial Intelligence Lab

ISCB ISMB 2022 MADISON

Introduction

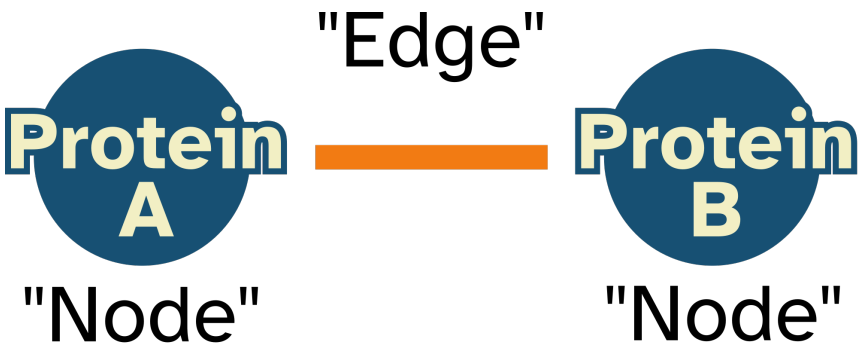
- Joseph Szymborski
  - McGill University,
    Department of Electrical & Computer Engineering
  - Mila, Quebec AI Institute
  - PhD Student in Amin Emad's COMBINE Lab

Emad's COMBINE Lab

Background: Protein-Protein Interactions

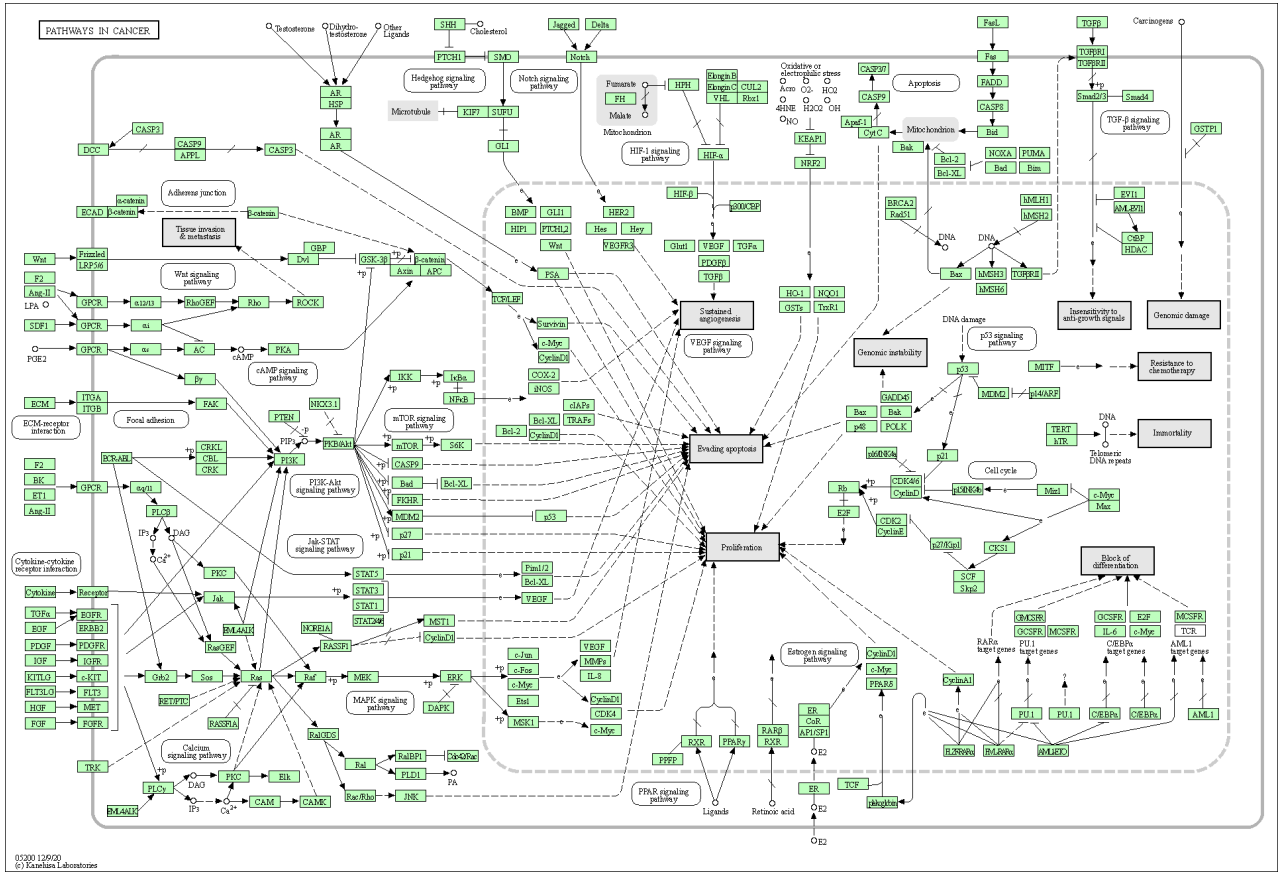
- Bio'l processes as an undirected graph of PPIs.

* An incomplete model,
but it's gotten us pretty far.



Emad's COMBINE Lab

Background: Protein-Protein Interactions

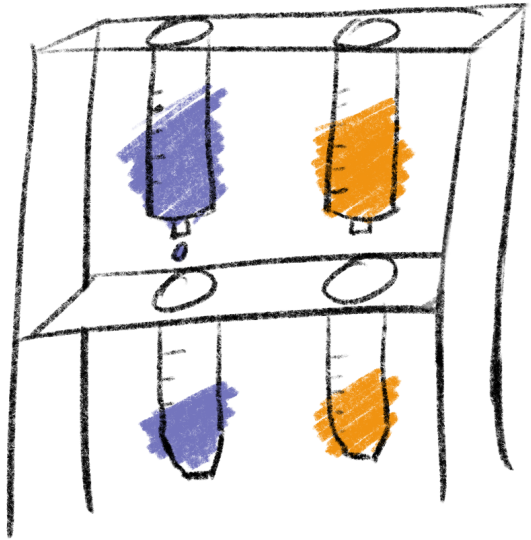


See:
Kanehisa M. *et al.*
10.1093/nar/gkr988

Emad's COMBINE Lab

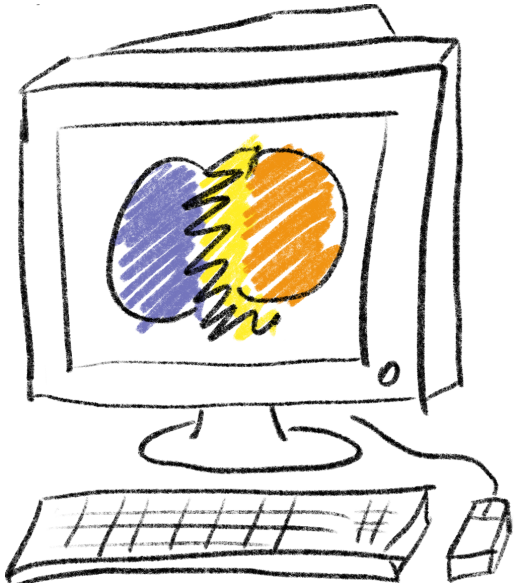
Background: Protein-Protein Interactions

- Protein interactions are typically identified through **“wet lab” experiments**.
- These experiments typically:
  - Take days/weeks.
  - Expensive reagents.
  - Often produce a lot of plastic waste.
  - Are quite definitive.



Background: Protein-Protein Interactions

- Predicting protein interactions using **computational models** try to address some of the **trade-offs of lab experiments**.
  - Take seconds/minutes.
  - Low-to-no cost.
  - Consume electricity and produces e-waste.
  - Not yet definitive.

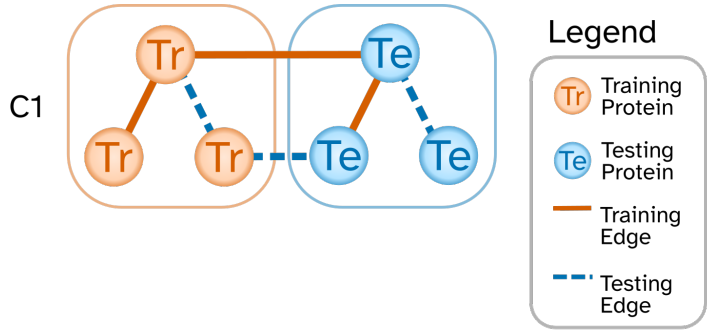


Background: Protein-Protein Interactions

Given two proteins, do they interact?

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

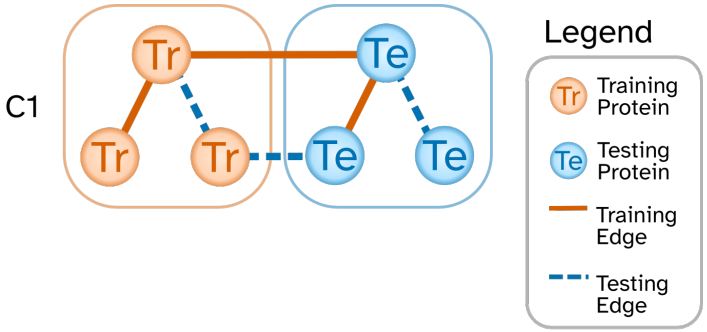


Park, Y. & Marcotte, E. M. (2012).

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

C1
0.81±0.01
0.85±0.01
0.64±0.01
0.77±0.01
0.81±0.01
0.77±0.01
0.56±0.01

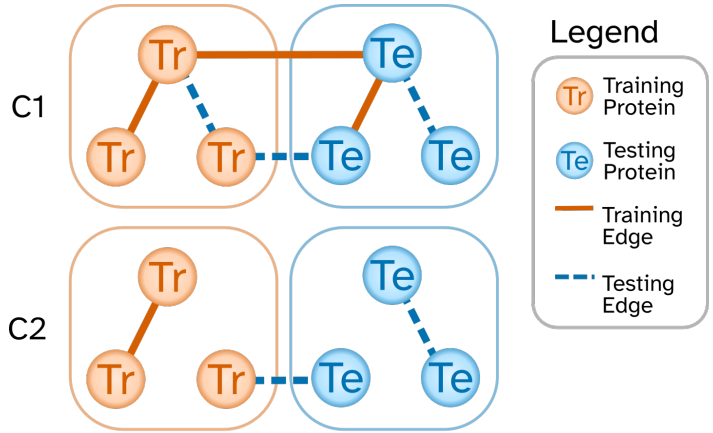


Park, Y. & Marcotte, E. M. (2012).

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

C1
0.81±0.01
0.85±0.01
0.64±0.01
0.77±0.01
0.81±0.01
0.77±0.01
0.56±0.01

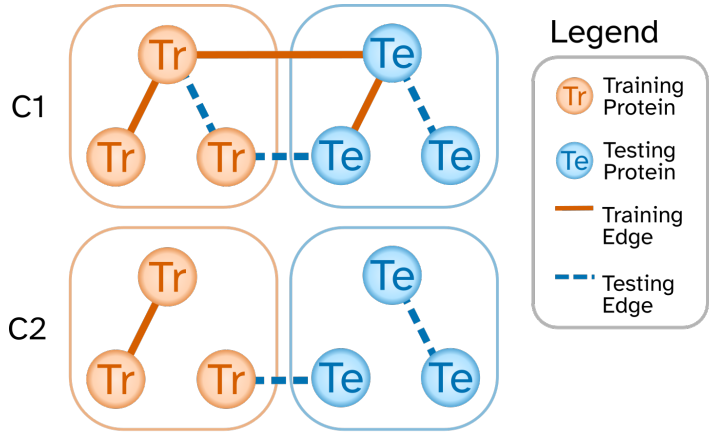


Park, Y. & Marcotte, E. M. (2012).

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

C1	C2
0.81±0.01	0.61±0.01
0.85±0.01	0.60±0.01
0.64±0.01	0.55±0.01
0.77±0.01	0.57±0.02
0.81±0.01	0.59±0.01
0.77±0.01	0.64±0.01
0.56±0.01	0.53±0.01

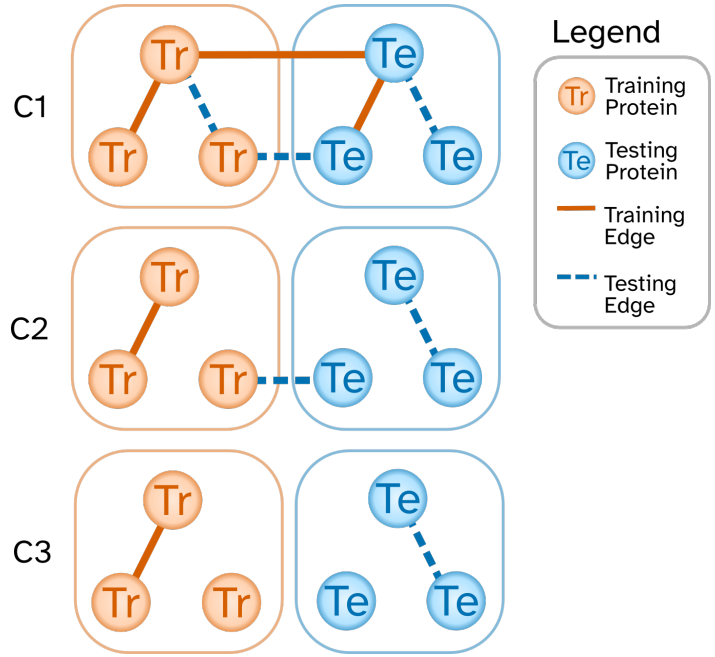


Park, Y. & Marcotte, E. M. (2012).

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

C1	C2
0.81±0.01	0.61±0.01
0.85±0.01	0.60±0.01
0.64±0.01	0.55±0.01
0.77±0.01	0.57±0.02
0.81±0.01	0.59±0.01
0.77±0.01	0.64±0.01
0.56±0.01	0.53±0.01

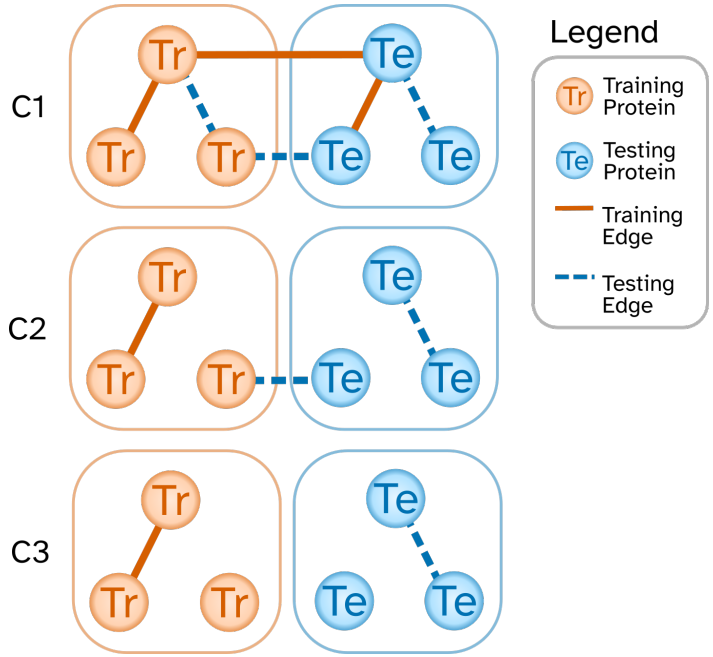


Park, Y. & Marcotte, E. M. (2012).

Background: Protein-Protein Interactions

AUROC on *H. sapiens*

C1	C2	C3
0.81±0.01	0.61±0.01	0.58±0.03
0.85±0.01	0.60±0.01	0.58±0.02
0.64±0.01	0.55±0.01	0.50±0.00
0.77±0.01	0.57±0.02	0.53±0.02
0.81±0.01	0.59±0.01	0.54±0.02
0.77±0.01	0.64±0.01	0.59±0.02
0.56±0.01	0.53±0.01	0.54±0.02



Park, Y. & Marcotte, E. M. (2012).

The Problem?

- It's hard to plug data leaks in PPI datasets.
- Many models depend on these leaks for their performance.
- How do we plug the leak?

Introducing 

Regularised **A**utomatic **P**rediction of **P**rotein-**P**rotein **I**nteractions using **D**eep Learning

Joseph Szymborski, Amin Emad, RAPPPID: Towards Generalisable Protein Interaction Prediction with AWD-LSTM Twin Networks, Bioinformatics, 2022; btac429, https://doi.org/10.1093/bioinformatics/btac429

Emad's COMBINE Lab

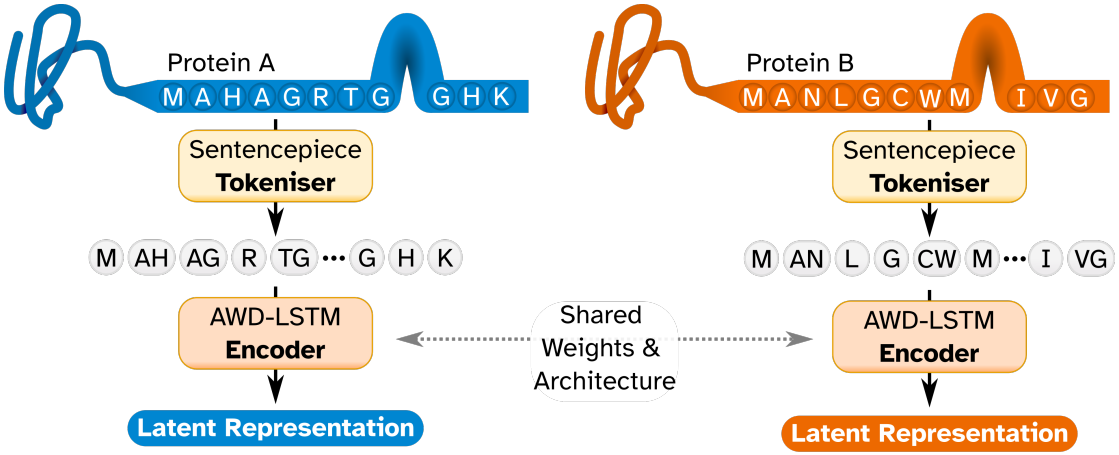
RAPPPID Architecture

Emad's COMBINE Lab

RAPPPID Architecture

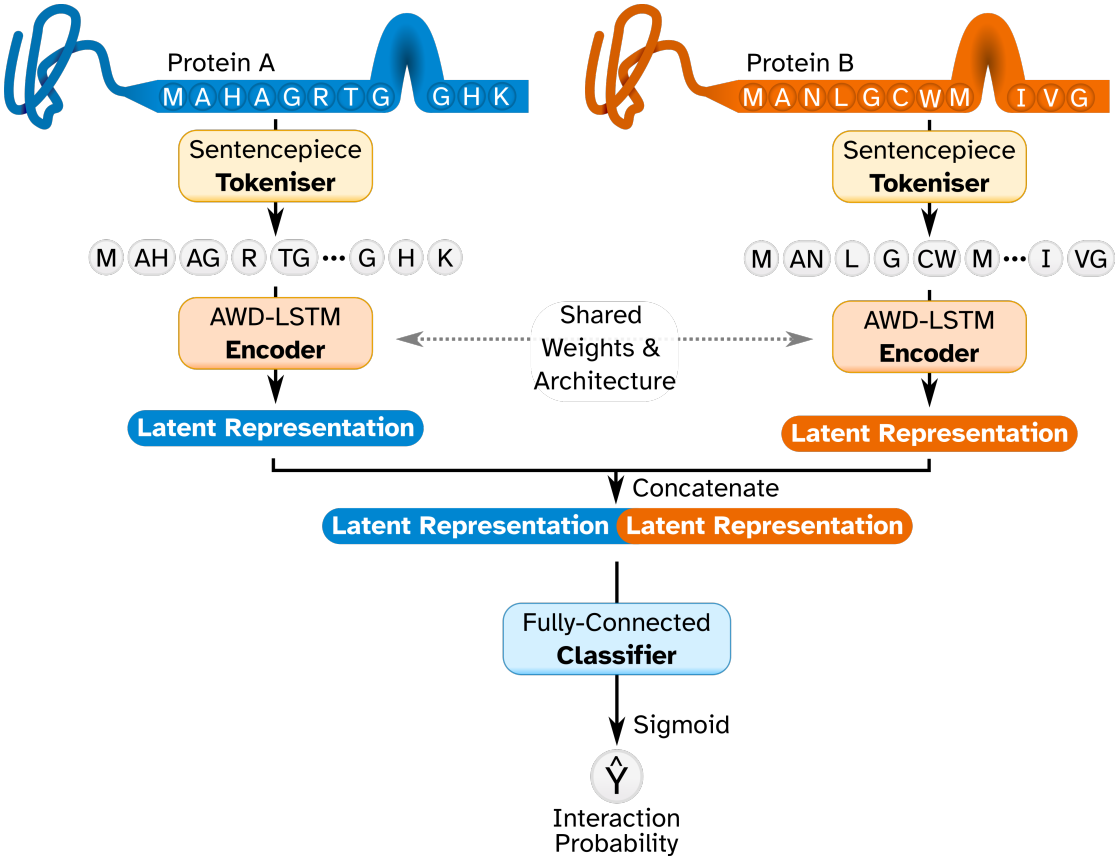
Emad's COMBINE Lab

RAPPPID Architecture



Emad's COMBINE Lab

RAPPPID Architecture



Emad's COMBINE Lab

What makes RAPPPID different?

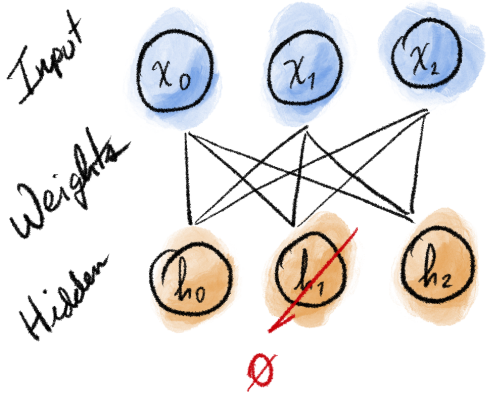
- In short, lots of regularisation
  - AWD-LSTM
  - Embedding dropout
  - Ranger21 Optimiser
  - Stochastic Weight Averaging (SWA)

What makes RAPPPID different?

- In short, lots of regularisation
  - AWD-LSTM
  - Embedding dropout
  - Ranger21 Optimiser
  - Stochastic Weight Averaging (SWA)
- Also
  - Sentencepiece tokenisation

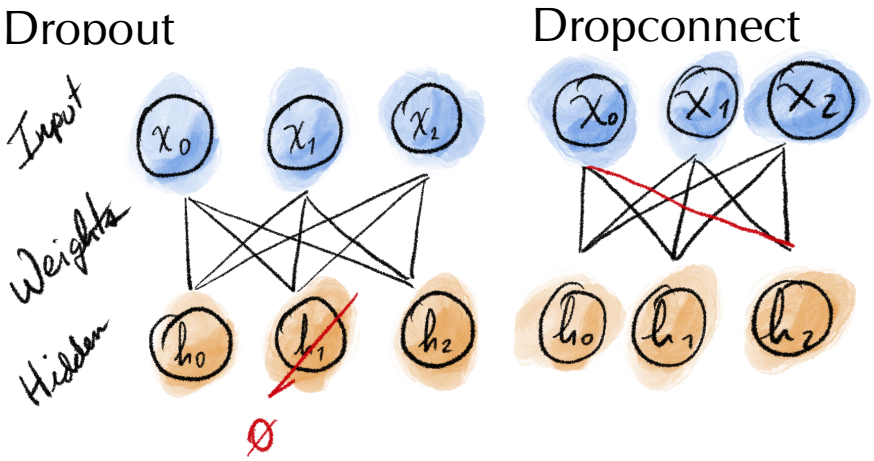
Regularising Recurrent Networks

Dropout



 Merity, S. *et al.* (2017)

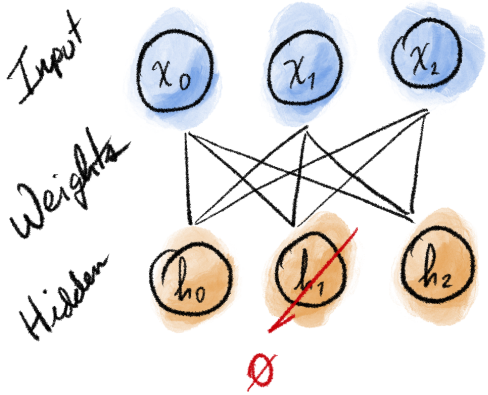
Regularising Recurrent Networks



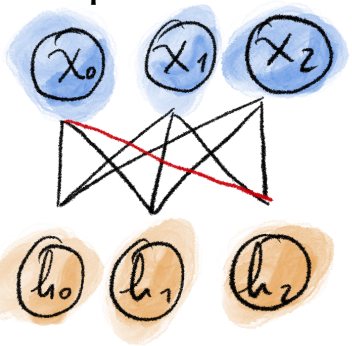
 Merity, S. *et al.* (2017)

Regularising Recurrent Networks

Dropout



Dropconnect



Embedding Dropout

```
A   1 0 0 ... 0 0 0
C   0 1 0 ... 0 0 0
D   0 0 1 ... 0 0 0
.         .
.         .
.         .
V   0 0 0 ... 1 0 0
W   0 0 0 ... 0 1 0 › 0 0 0 ... 0 0 0
Y   0 0 0 ... 0 0 1
```

 Merity, S. *et al.* (2017)

Stochastic Weight Averaging (SWA)

- Very similar to Averaged Stochastic Gradient Descent but keeps a pair of weights:
  - One that the optimiser minimises (w).
  - Another that is a running average of the previous weight (w_{SWA}).

$$w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{models} + w}{n_{models} + 1}$$

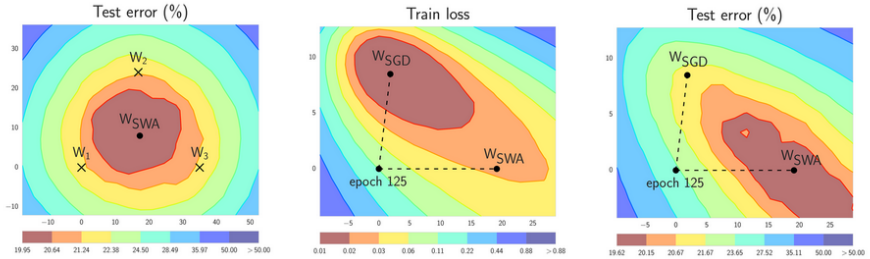
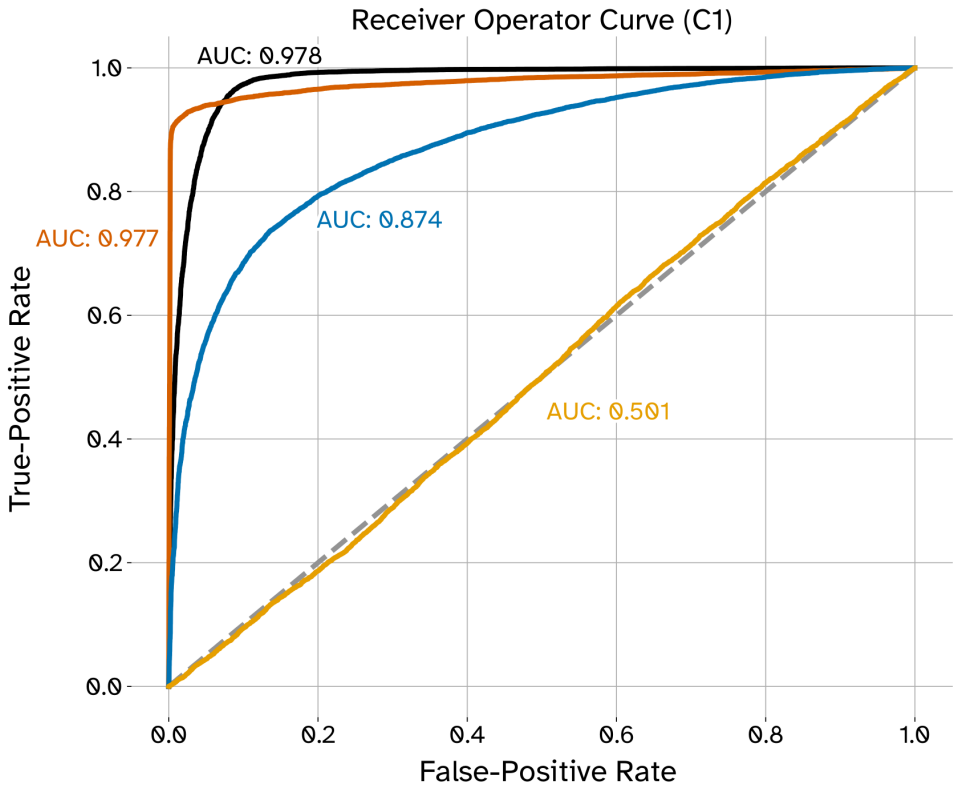
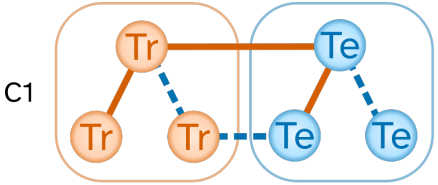
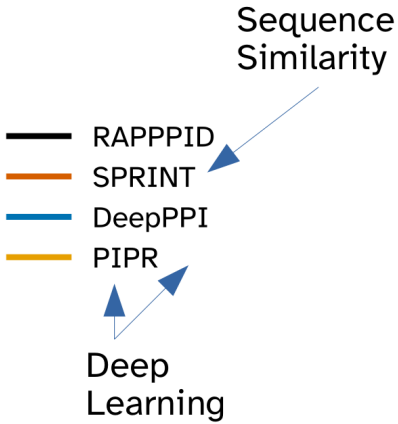


Figure 1. *Illustrations of SWA and SGD with a Preactivation ResNet-164 on CIFAR-100 [1]. **Left**: test error surface for three FGE samples and the corresponding SWA solution (averaging in weight space). **Middle** and **Right**: test error and train loss surfaces showing the weights proposed by SGD (at convergence) and SWA, starting from the same initialization of SGD after 125 training epochs. Please see [1] for details on how these figures were constructed.*

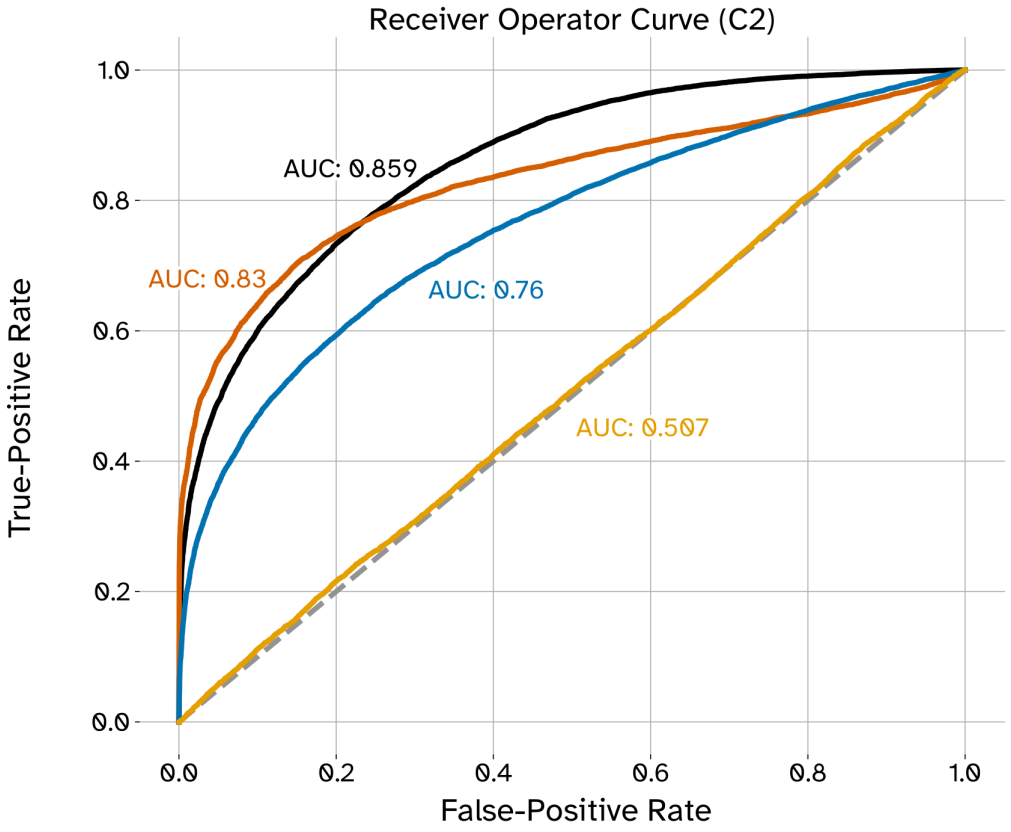
How does RAPPPID perform?

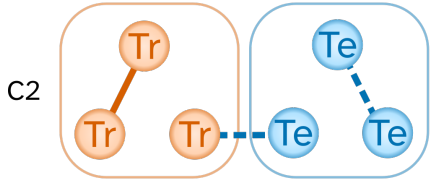






How does RAPPPID perform?





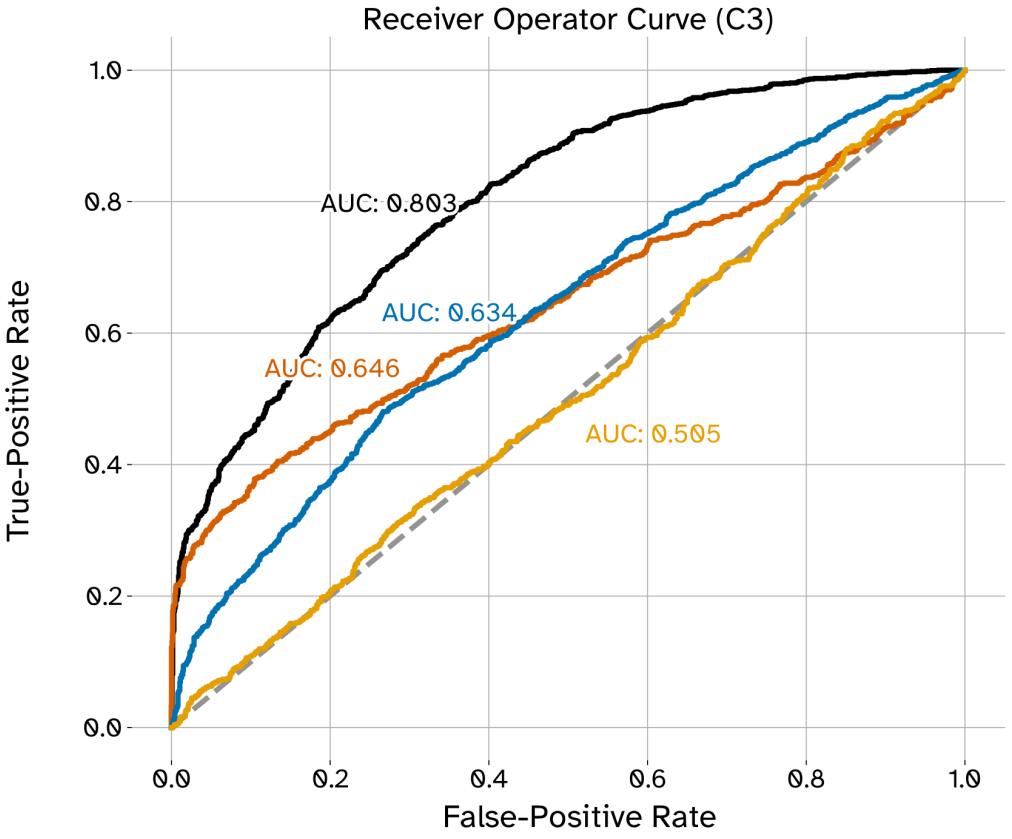
Sequence Similarity

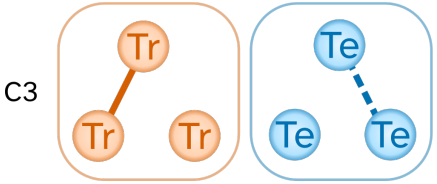
RAPPPID
SPRINT
DeepPPI
PIPR

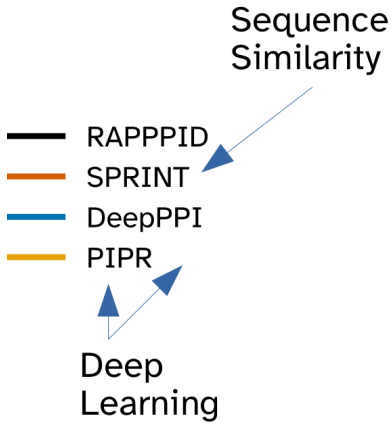
Deep Learning

Emad's COMBINE Lab

How does RAPPPID perform?







Emad's COMBINE Lab

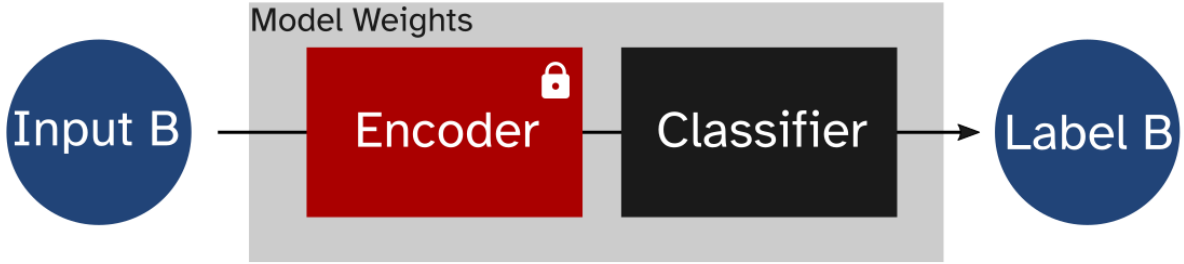
RAPPPID performance vs. data providence

Results from an ablation study conducted on RAPPPID. Each model is trained/tested twice on three randomly generated C3 datasets. The performance metrics correspond to held-out test sets.

	RAPPPID (original)	RAPPPID -SWA	RAPPPID + Adam	RAPPPID- AWD	RAPPPID- SentencePiece	RAPPPID + TransfLG	RAPPPID + TransfSM
Test AUROC	0.792 (±0.007)	0.782 (±0.007)	0.791 (±0.025)	0.762 (±0.020)	0.749 (±0.009)	0.670 (±0.030)	0.747 (±0.026)
AUROC Diff	N/A	-1.20%	-0.100%	-3.70%	-5.37%	-15.3%	-5.68%
Test APR	0.794 (±0.009)	0.783 (±0.007)	0.792 (±0.032)	0.757 (±0.022)	0.748 (±0.011)	0.686 (±0.040)	0.758 (±0.025)
APR Diff	N/A	-1.37%	-0.273%	-4.62%	-5.85%	-13.6%	-4.61%

Transfer Learning on X-Ray Crystallography Data

- BioLIP dataset: semi-curated dataset of Protein/Ligand interactions based on the PDB
- We pretrain on STRINGDB, then fine-tune on BioLIP

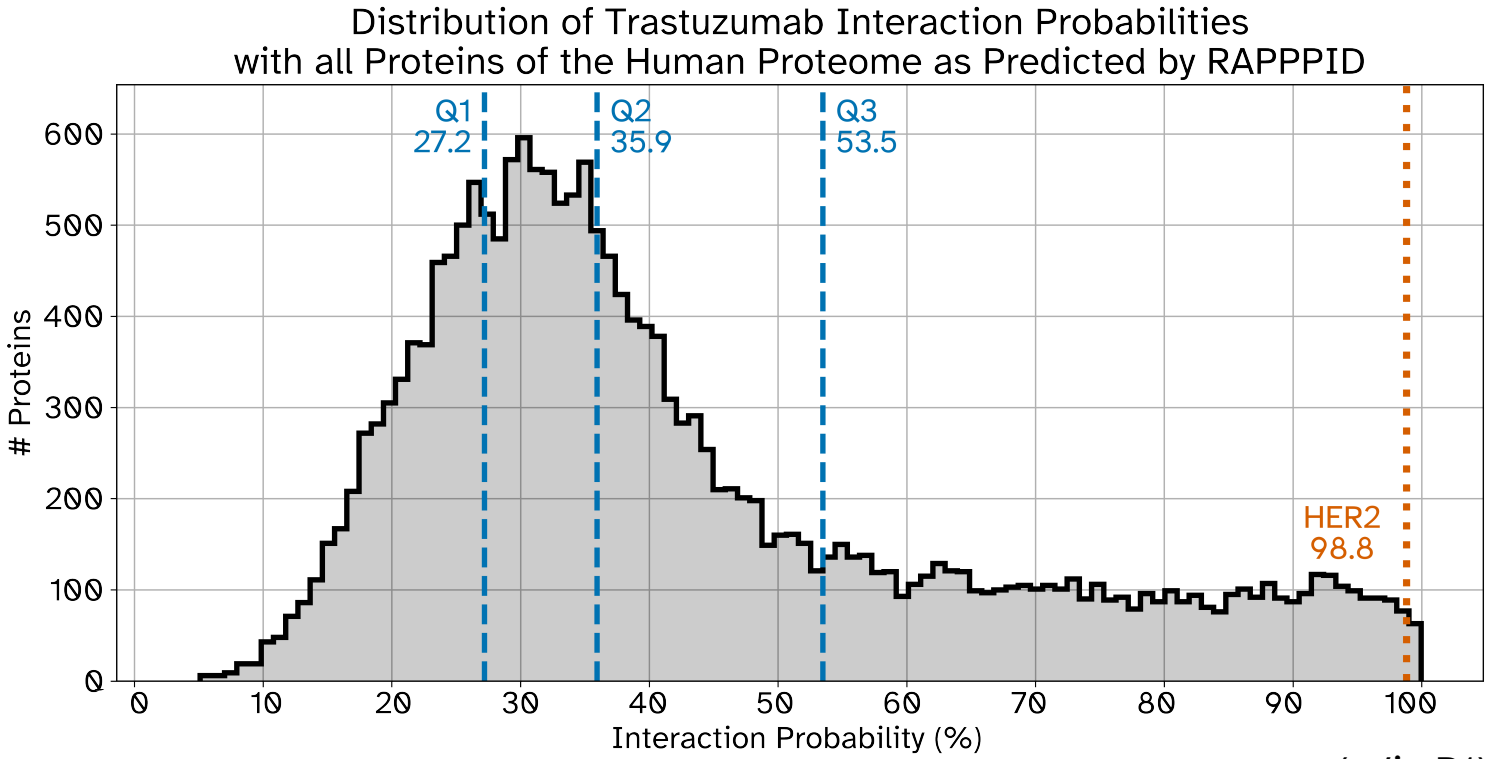


- Training on STRING DB, fine-tuning on BioLIP, and testing on BioLIP:
  - AUROC of **0.909**

RAPPPID predicts interaction of HER2 with Trastuzumab and Pertuzumab

- How might one use RAPPPID to validate hypothesized interactions between:
  - Target proteins
  - Candidate therapeutic proteins and peptides
- Two examples: Trastuzumab and Pertuzumab.
  - Recombinant humanised monoclonal antibodies
  - Used for HER2-positive metastatic breast cancer

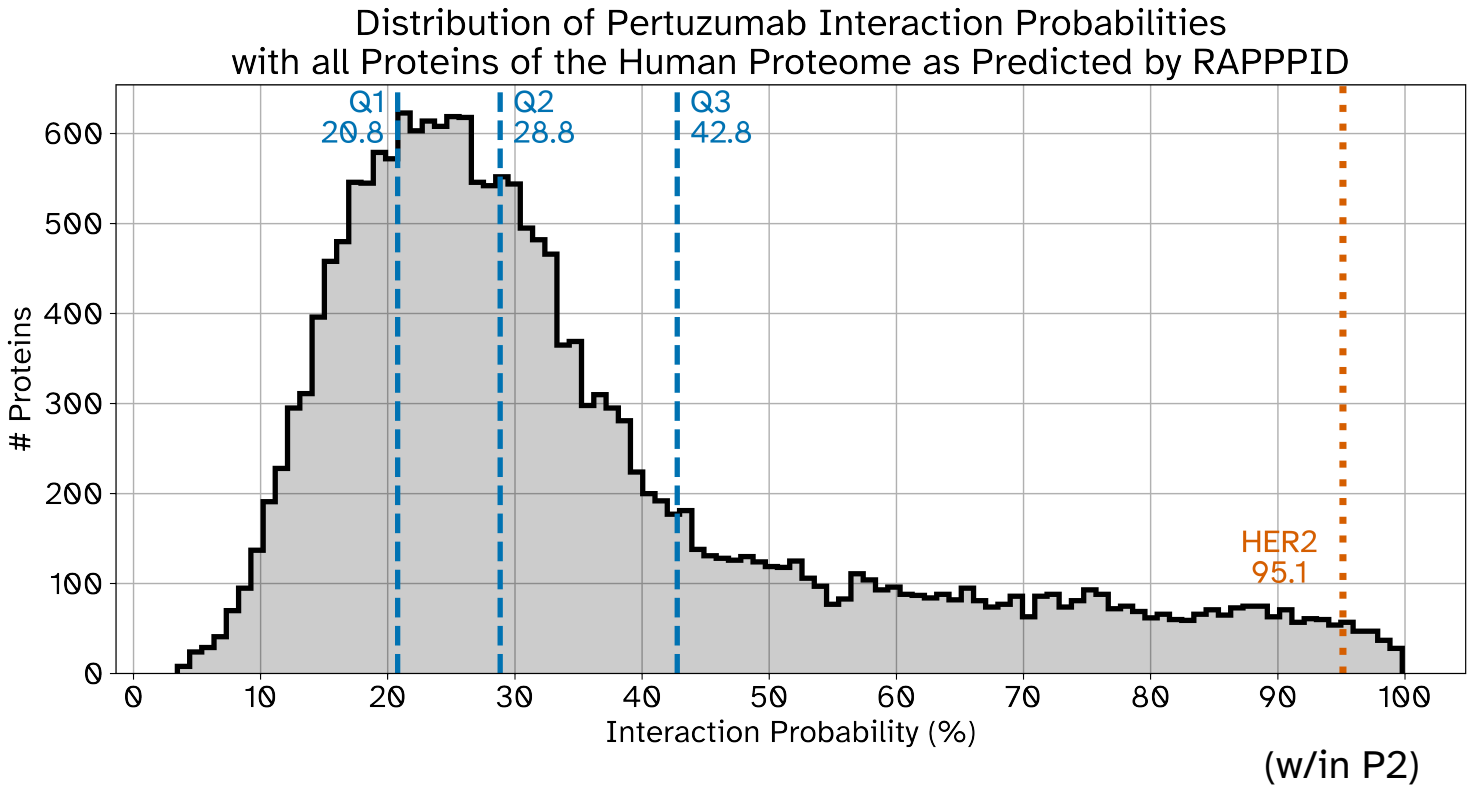
RAPPPID predicts interaction of HER2 with Trastuzumab and Pertuzumab



(w/in P1)

Emad's COMBINE Lab

RAPPPID predicts interaction of HER2 with Trastuzumab and Pertuzumab



Emad's COMBINE Lab

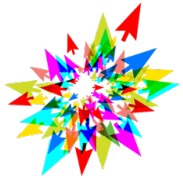
Acknowledgements

- Thanks to the members of the COMBINE lab for their feedback and support
  - P.D.F: Antoine Soulé
  - Ph.D.: David E. Hostallero, Ali Saberi, Yitian Zhang
  - M.Sc.: Mohamed Reda El Khili, Jessica (Yihui) Li, Chen Su, Abulrahman Takiddeen

Thanks to our supporters:

VADASZ SCHOLARS

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION



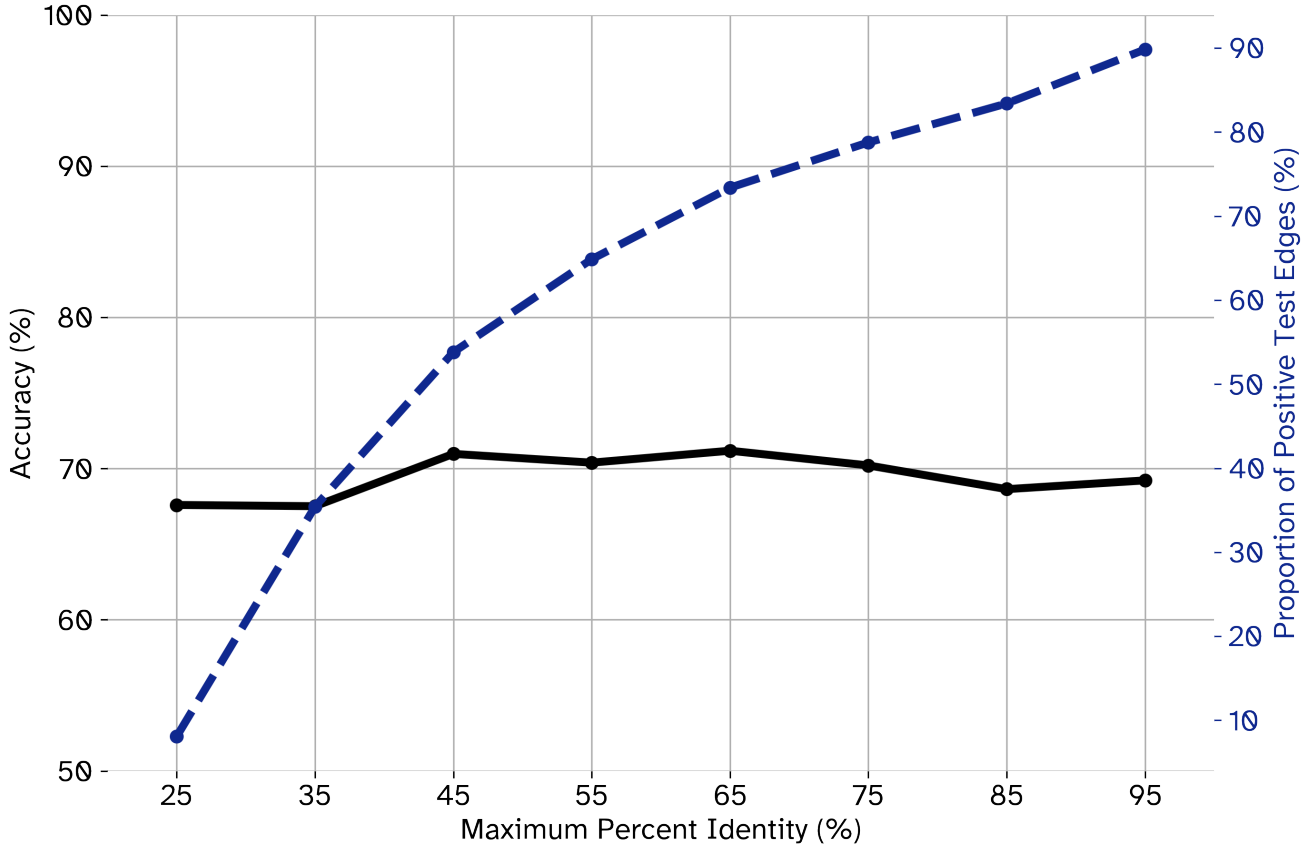
compute canada | calcul canada

Calcul Québec

Thank you

Questions?

Is RAPPPID just identifying similar sequences?



A1

Emad's COMBINE Lab

Existing PPI datasets are not great for Deep Learning.

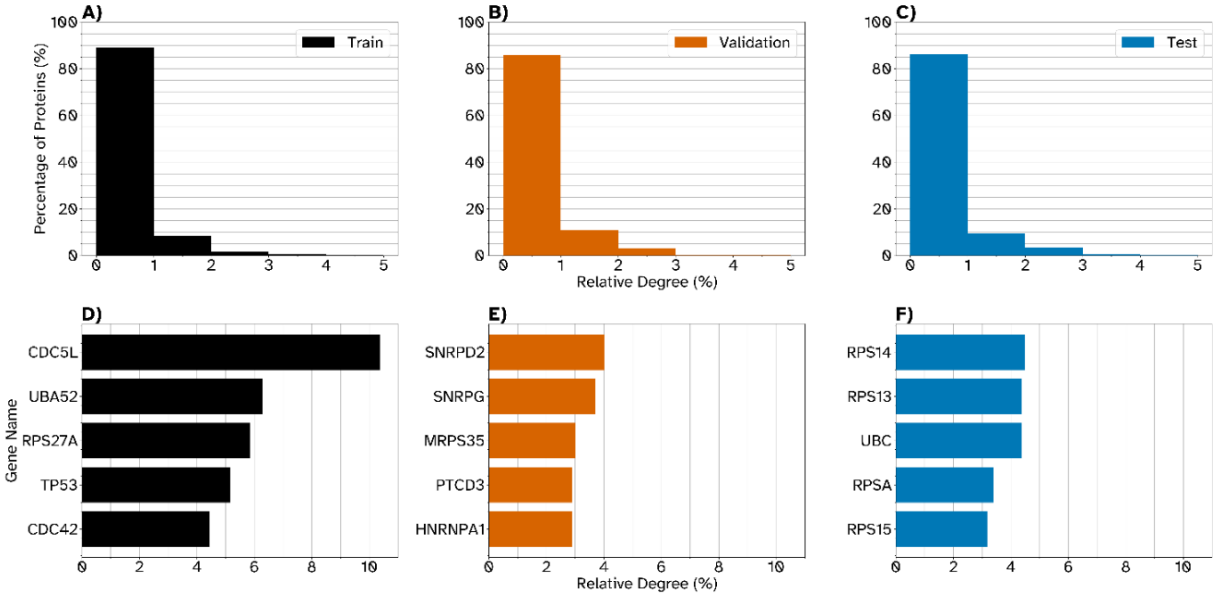
- We wanted to use additional datasets, like HIPPIE and iRefWeb
- Only STRING has enough high-confidence edges for deep learning purposes
  - 98.5% fewer edges in HIPPIE than in STRING (human, 95% confidence)
  - 87.9% fewer edges with an 85% confidence.
  - 75% fewer edges in iRefWeb than in STRING (human, 95% confidence)
- This is made worse by the fact that PPI datasets overfit terribly to begin with

False-Positive Rate

- We evaluated the false-positive rate of confidence score-filtered STRING dataset
  - We used curated and experimentally validated non-interacting protein pairs from **Negatome**
- We compared the set of proteins that are:
  - Both in STRING and Negatome
  - Evaluating the number of negative edges in Negatome that were considered a positive edge in this interesection
- Estimated the false-positive rate of our STRING dataset to be **4.01%**
- Falls within the extected **5%** upper-bound given by our 95% confidence threshold

Protein Over-Representation

- PPI graphs are understood to be scale-free in the general case
- That means that some hub proteins might be over-represented
- But that isn't the case.



Emad's COMBINE Lab

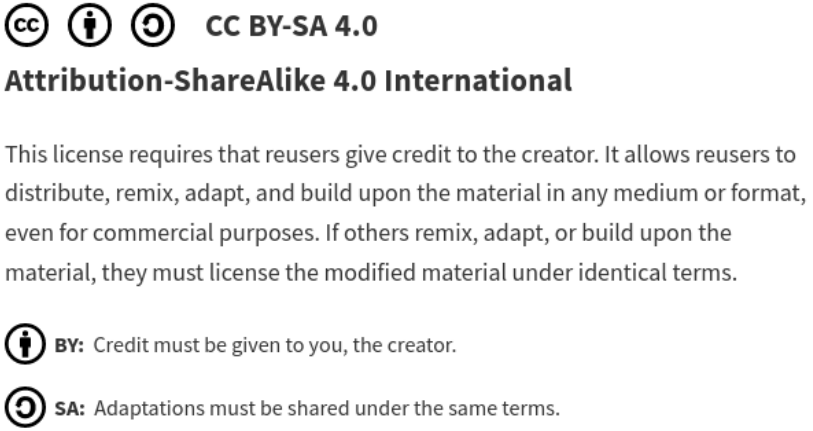
Curated negative examples

- We investigated using the curated database “Negatome” for the negative samples
- There are too few (1,191 negative *H. sapiens* pairs; 263,130 positive pairs)

License

- These slides are licensed as follows:

  BIRS 2022 Presentation: RAPPPID: Improving Protein Interaction Prediction on Unseen Proteins © 2022 by Joseph Szymborski is licensed under Attribution-ShareAlike 4.0 International. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/



Emad's COMBINE Lab