SCLpred: *Protein subcellular localization prediction by deep N-to-1 neural networks*

Liam Ellinger[1,2,3], Manaz Kaleel[1], Clodagh Lalor[1], Gianluca Pollastri[1] & Catherine Mooney[1,*]

[1]**School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland;**

[2]**College of Global Studies, Arcadia University, Glenside, PA, USA;**

[3]**Whitacre College of Engineering, Texas Tech Universiy, Lubbock, TX, USA**

[*]catherine.mooney@ucd.ie

1. Introduction

The subcellular localization of a protein is a crucial piece of information in order to tell how the protein functions in the body. In particular, proteins that make up and reside in/on the cell's membranes are interesting because the cell uses these as signals, receptors, identification and transport of molecules in and out of the cell. These are the proteins that are the most interacted with by other cells and by extension the body as a whole. This project aims to create a predictor of the location of a protein in a cell, specifically if the protein will be sent to or "targeted" to a membrane after it has been synthesised. The predictor uses a convolutional neural network trained in 5-fold cross-validation on a large 80% redundancy reduced dataset consisting of all proteins added to Uniprot [1] that have experimentally verified subcellular localizations added after 2016. Each protein has its sequence plus its evolutionary information encoded via PSI-BLAST[2] profiles. By testing SCLPred and other predictors with the same independent test set, we show that SCLPred is comparable if not superior to other state of the art predictors.

2. Materials and Methods

Datasets

- UniProt release 2019_05 [1]
- All entries from humans, mammals, invertebrates, plants, fungi
- 189,818 protein sequences from 7,859 species
- Initially redundancy reduced to less than 80% sequence similarity, then further reduced internally using a e-value of 0.001 in PSI-BLAST
- Split into training set (TS) and independent test set (ITS) – sequences that were added to UniprotKB after 2016

Hyperparameter Tuning

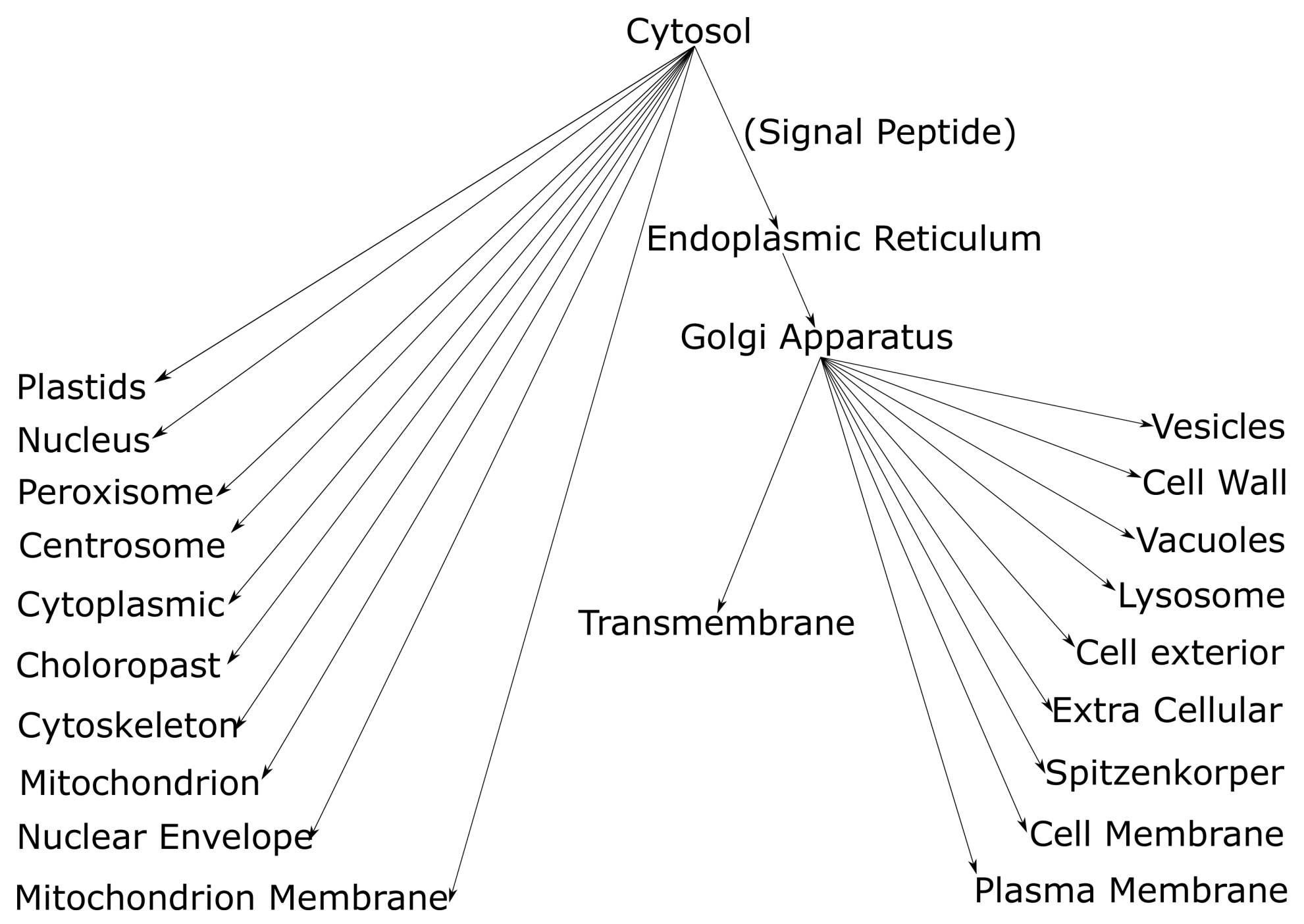
- Network contains tunable parameters customizable to the best general performance
- Many different configurations tested: number of kernel layers, width of their inputs and outputs, with of data examined with each step and others
- The best performing configuration in our testing became the configuration used in 5-fold cross-validation

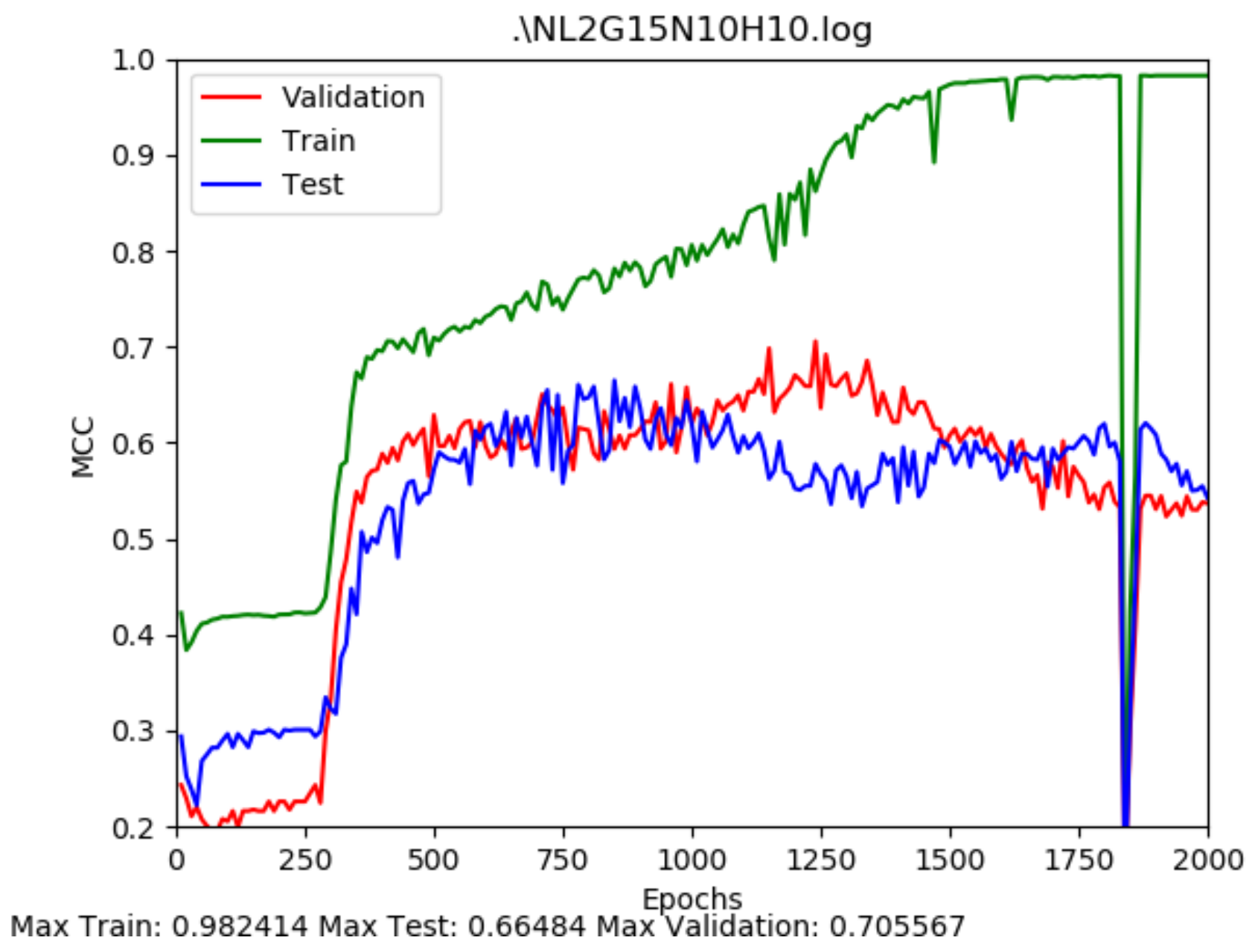
Predictive architecture

- The prediction algorithm employs a modified Convolutional Neural Network (CNN)
- Trained and tested in 5-fold cross-validation
- We use Matthews correlation coefficient (MCC) as a measure of the correlation between observed and predicted states

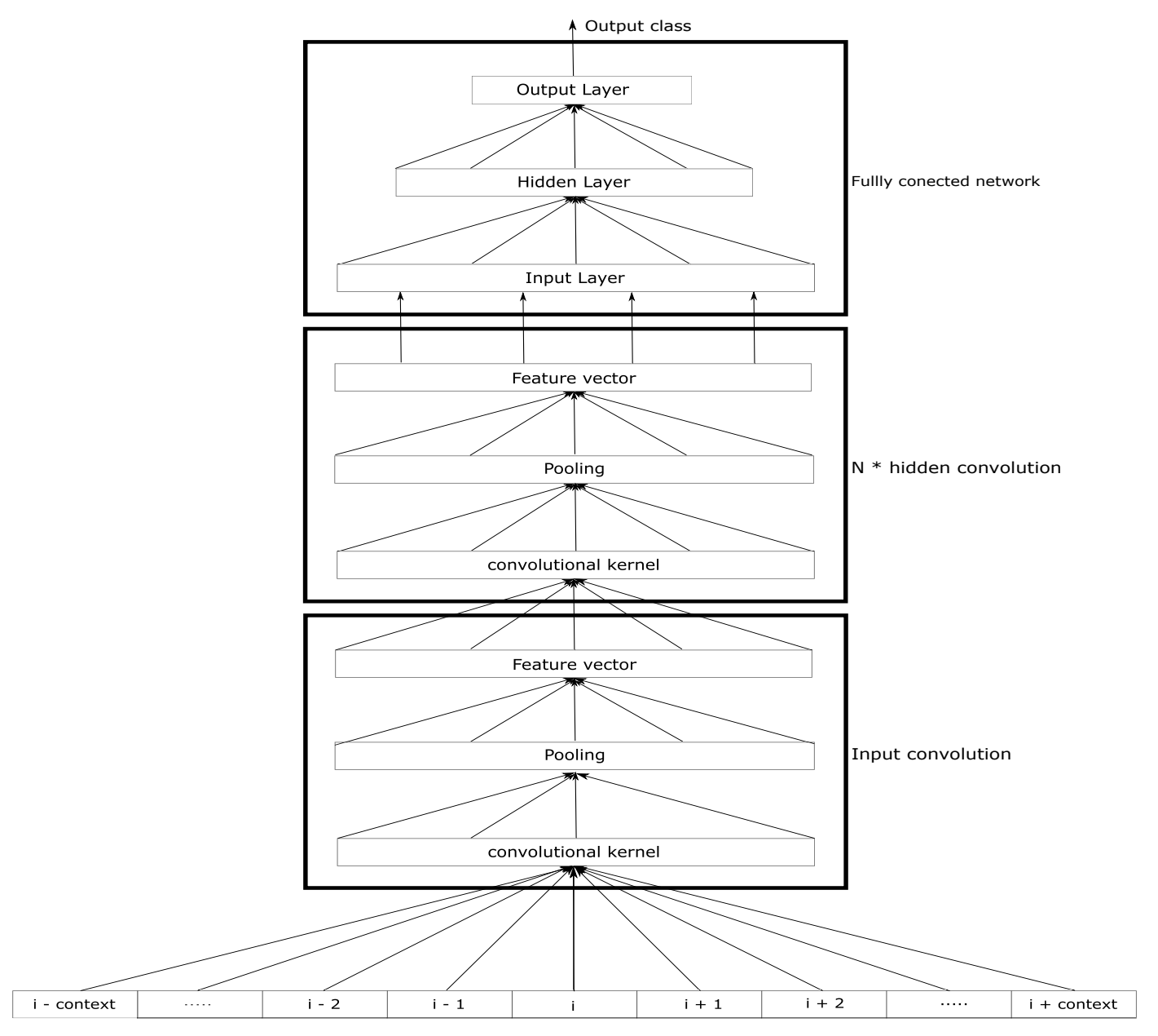
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

	TS	ITS
Mem	2,766	45
Other	2,026	73
Total	4,792	118





Graph of the performance of the best configuration over 2000 epochs

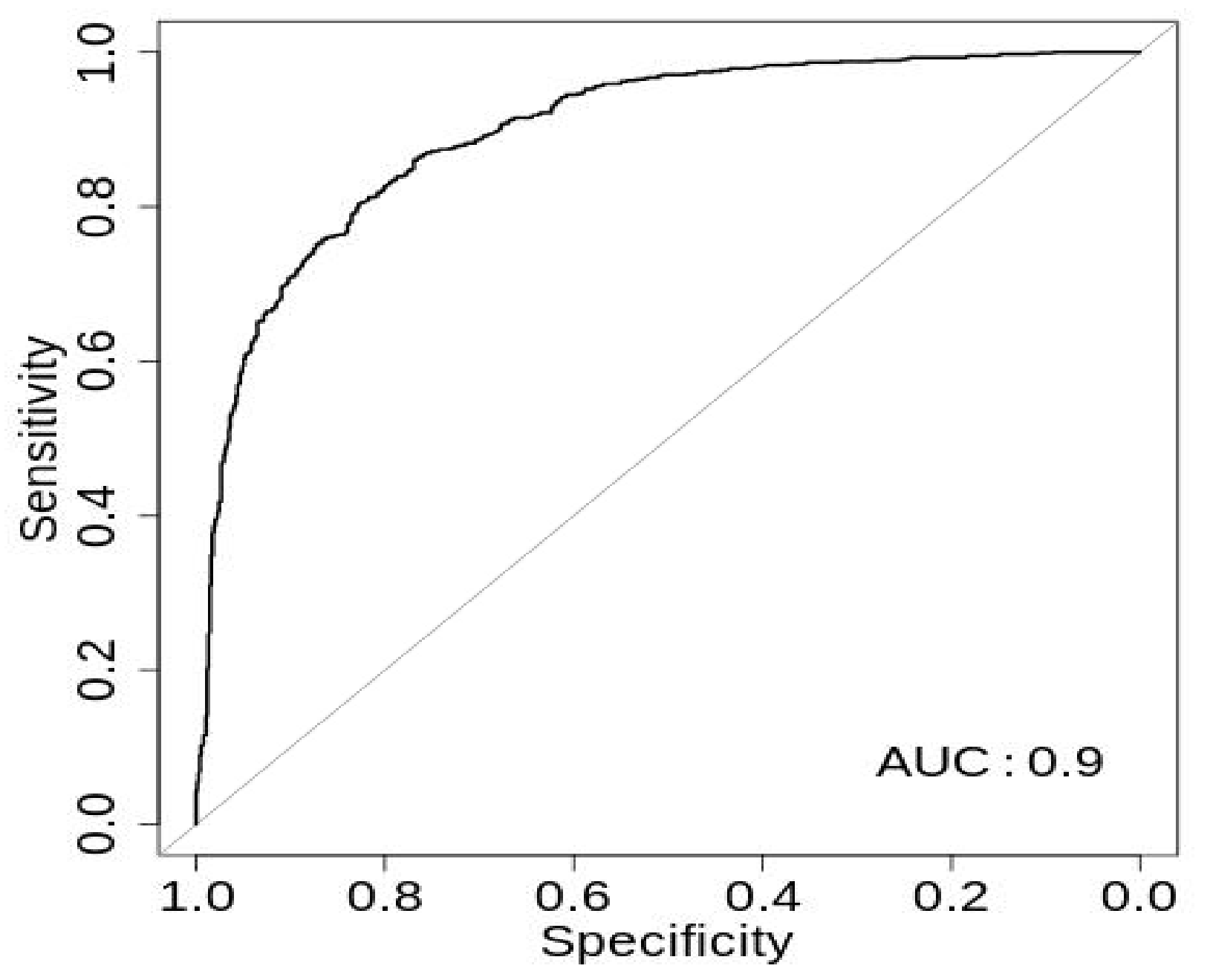


Neural Network Architecture

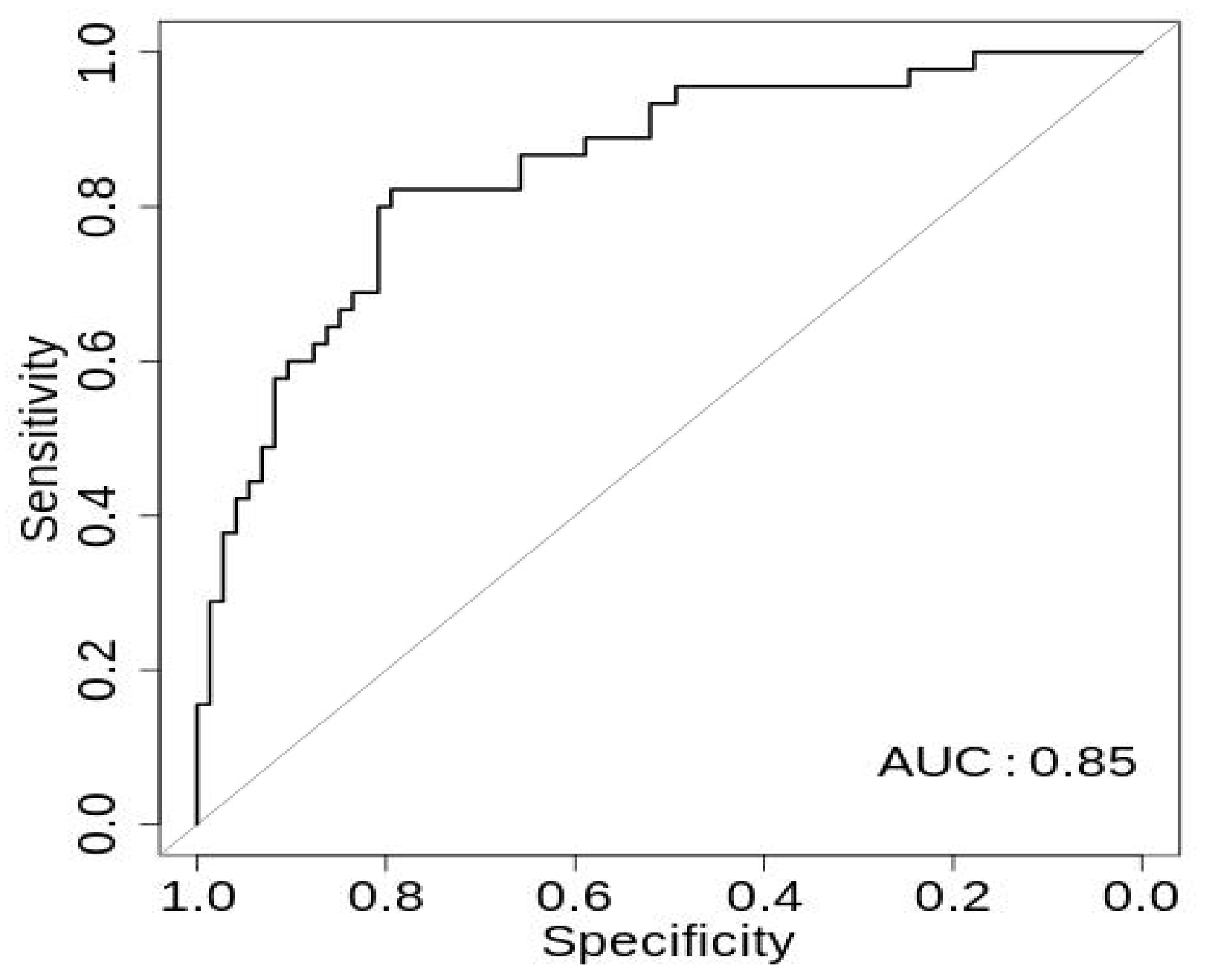
3. Results

Treatments	MCC	Accuracy
DeepLoc [3]	0.227	66.1%
SCLPred (ITS)	0.521	78.0%
SCLPred (TS)	0.625	81.2%

The performance of SCLpred compared to DeepLoc [3] on the ITS. Left chart is a ROC chart showing SCLPred's test set, right is ROC chart showing ITS.



SCLpred (TS)



SCLpred (ITS)

4. Conclusions

- SCLpred is a state-of-the-art protein subcellular localization prediction tool
- We predict subcellular localization into two classes: membranes and everything else
- SCLpred achieves an MCC of 0.52 on an independent test set of 118 protein sequences added to UniProt since 2016

5. Future Work

Current work is focused on this specific task, but this predictor may be used as a part of a larger, more general predictor predicting > 2 classes. This predictor could also be made public via a webserver.

6. References

[1] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[3] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

7. Acknowledgements

The work of M.K. is supported by a grant from the Irish Research Council [GOIPG/2014/603].
The authors acknowledge the Research IT Service at University College Dublin for providing HPC resources that have contributed to the research results reported within this poster. `http://www.ucd.ie/itservices/ourservices/researchit/`