Graph neural networks for investigating complex diseases: A case study on Parkinson's Disease

Elisa Gómez de Lope[1], Ramón Viñas Torné[2], Pietro Liò[2], Enrico Glaab[1] on behalf of the NCER-PD consortium

[1]Biomedical Data Science Group, LCSB, University of Luxembourg
[2]Department of computer science, University of Cambridge

@elisagdelope
elisa.gomezdelope@uni.lu

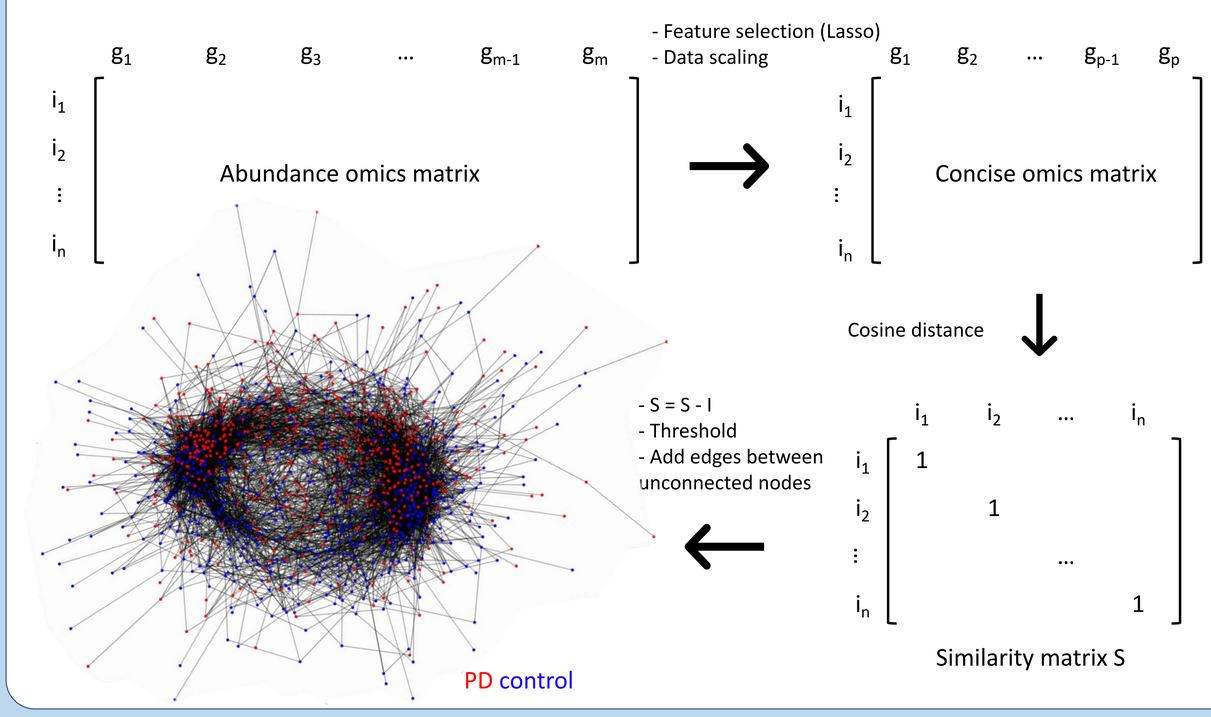
DOWNLOAD POSTER HERE!

Background

Graph neural networks (GNNs) have emerged as a promising approach to investigate relational information. Omics data analysis is a critical component in the study of complex diseases, and allows to represent relational information among samples as a graph structure that can be modelled with GNNs. However, it is still unclear which strategies for designing and optimizing GNNs are most effective when working with real-world omics data from complex disorders, such as Parkinson's disease (PD).

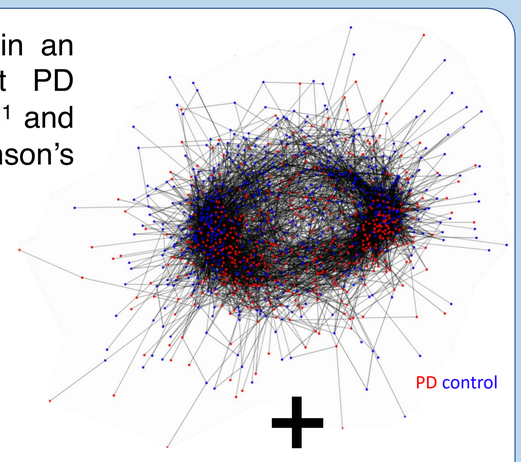
Methods

Here we examined various GNN models to identify and interpret discriminative patterns between PD patients and controls using omics data. We built a pipeline integrating 1) Lasso penalty-based feature selection; 2) similarity graph construction based on cosine distance; 3) modelling for sample (node) classification.

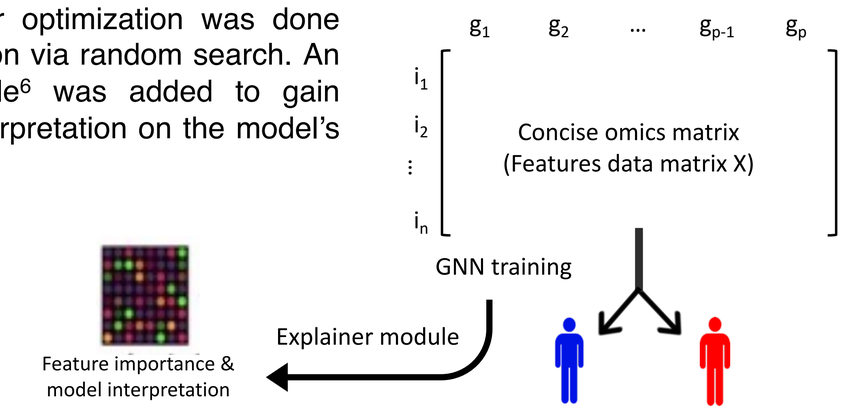


The pipeline was trained and evaluated in an end-to-end manner in two independent PD omics datasets (transcriptomics from PPMI[1] and metabolomics from the Luxembourg Parkinson's Study[2]) with models:

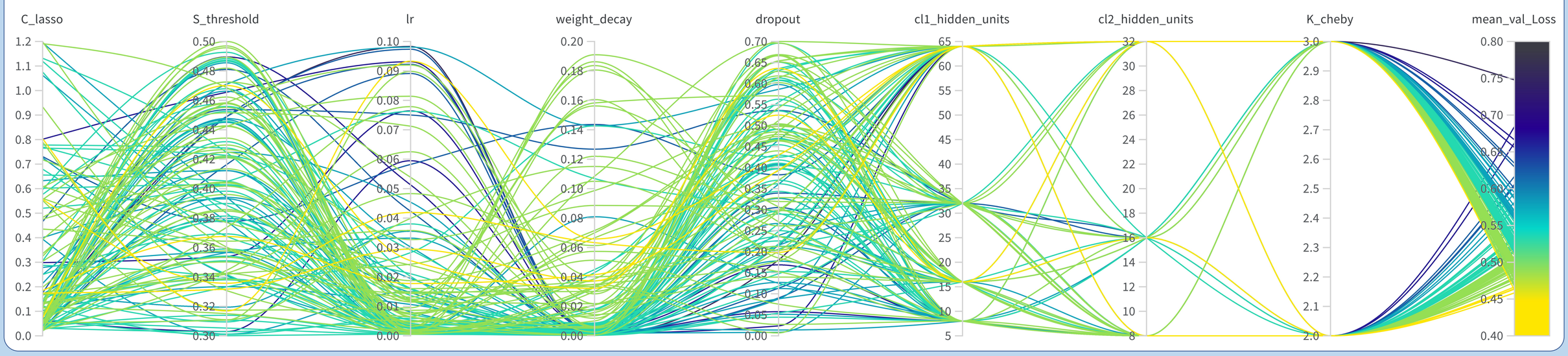
- Graph Convolutional Network[3]
- ChebyNet[4]
- Graph Attention Network[5]



Hyper-parameter optimization was done in cross validation via random search. An explainer module[6] was added to gain insights and interpretation on the model's decisions.



An extensive random hyperparameter search was performed for the validation set; in each fold 130 runs were launched exploring values of regularization penalty, similarity (edge) threshold, learning rate, weight decay, dropout, number of convolutional layer units, and K (for ChebyNet model). Despite some variability, certain trends are visible: the best models (i.e., with lower average validation loss) tend to be achieved when avoiding higher learning rates in combination with low weight decay and low dropout.



Results

Test performance of GCN, ChebyNet, GAT in transcriptomics dataset[5]

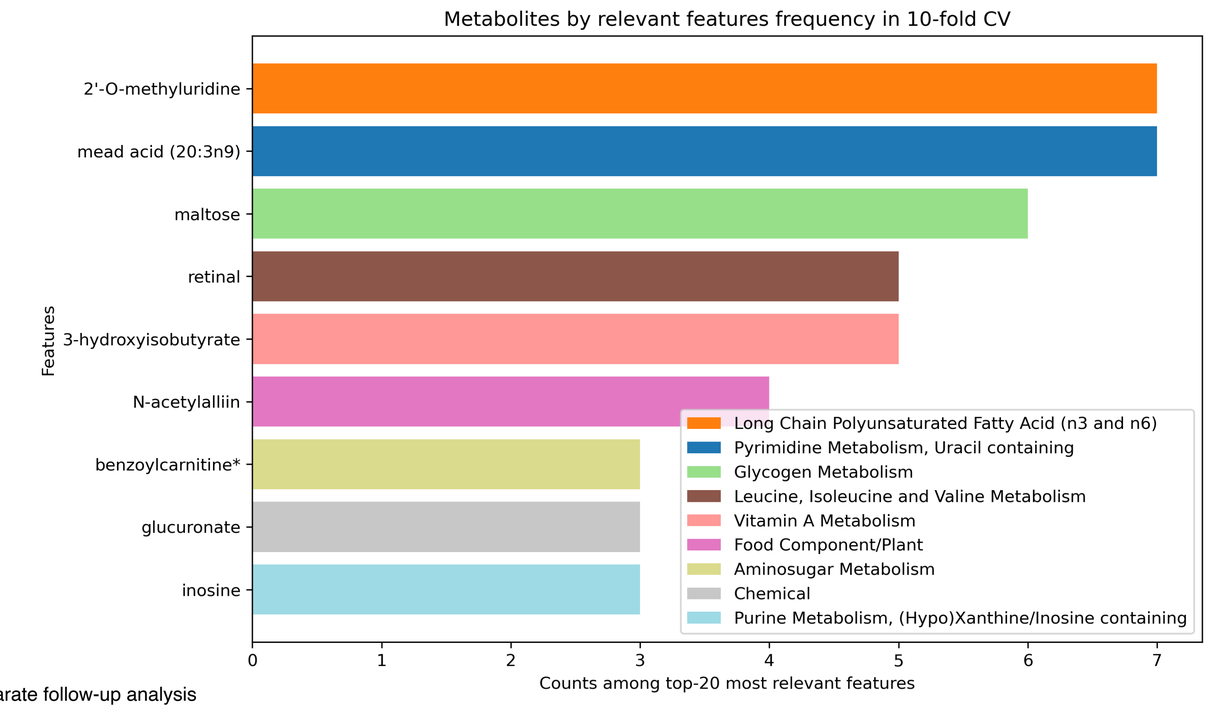
	AUC	Accuracy	Recall	Specificity
ChebyNet	0.55 ± 0.05	0.53 ± 0.05	0.58 ± 0.05	0.49 ± 0.12
ChebyNet(w)	0.58 ± 0.05	0.55 ± 0.06	0.55 ± 0.1	0.55 ± 0.08
GCN	0.55 ± 0.07	0.56 ± 0.07	0.55 ± 0.07	0.56 ± 0.11
GCN (w)	0.56 ± 0.04	0.53 ± 0.04	0.56 ± 0.07	0.51 ± 0.08
GAT	0.53 ± 0.05	0.51 ± 0.06	0.53 ± 0.08	0.49 ± 0.08
GAT (w)	0.56 ± 0.04	0.54 ± 0.04	0.59 ± 0.06	0.48 ± 0.08
SVM Radial (no graph)	0.64 ± 0.07	0.6 ± 0.06	0.58 ± 0.07	0.65 ± 0.09

Test performance of GCN, ChebyNet, GAT in metabolomics dataset*[6]

	AUC	Accuracy	Recall	Specificity
ChebyNet	0.83 ± 0.05	0.75 ± 0.05	0.77 ± 0.05	0.73 ± 0.12
ChebyNet (w)	0.83 ± 0.05	0.74 ± 0.06	0.72 ± 0.1	0.76 ± 0.08
GCN	0.78 ± 0.07	0.72 ± 0.07	0.7 ± 0.07	0.74 ± 0.11
GCN (w)	0.81 ± 0.04	0.74 ± 0.04	0.72 ± 0.07	0.75 ± 0.08
GAT	0.79 ± 0.05	0.74 ± 0.06	0.69 ± 0.08	0.78 ± 0.08
GAT (w)	0.79 ± 0.04	0.73 ± 0.04	0.69 ± 0.06	0.76 ± 0.08
SVM Radial (no graph)	0.88 ± 0.07	0.8 ± 0.06	0.8 ± 0.07	0.81 ± 0.09

(w) = weighted network was used in the model
*This is a comparison focusing on methodology; the metabolomics dataset contains treatment confounding effects requiring separate follow-up analysis

Metabolites that most frequently appeared among the top-20 most relevant for 10-fold CV ChebyNet model on a subset of unmedicated *de novo* PD patients vs controls from the metabolomics cohort



Conclusions

- In this implementation, the attention mechanism in GATs did not provide advantages when compared to GCN and ChebyNet, while ChebyNet performed better than GCN.

- Contrary to previous research on graph classification tasks, using a GCN layer did not beat the more established methods that only take a flatten representation into account (i.e. SVM).

- We conjecture that high levels of noise combined with limited sample size hinder graph convolutional operators from learning meaningful representations from single omics, hence learning similar embeddings regardless of the diagnosis. Incorporating molecular interactions data or multi-omics from the same cohort hold potential to capture richer node embeddings.

[1] Marek, K. et al. *The Parkinson Progression Marker Initiative (PPMI)*. Prog Neurobiol 95, 629 (2011).
[2] Hipp, G. et al. *The Luxembourg Parkinson's Study: A Comprehensive Approach for Stratification and Early Diagnosis*. Front Aging Neurosci 10, (2018).
[3] Thomas N. Kipf, Max Welling (2017). *Semi-Supervised Classification with Graph Convolutional Networks*. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings https://doi.org/10.48550/arXiv.1609.02907
[4] Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst (2016). *Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering*. 30th International Conference on Neural Information Processing Systems, 2016 - Conference Track Proceedings https://doi.org/10.48550/arXiv.1606.09375
[5] Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., & Bengio, Y. (2018). *Graph attention networks. 6th International Conference on Learning Representations*, ICLR 2018 - Conference Track Proceedings https://doi.org/10.17863/CAM.48429
[6] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec (2019). *GNNExplainer: Generating Explanations for Graph Neural Networks*. https://doi.org/10.48550/arXiv.1903.03894