BraneMF: Random Walk-based Matrix Factorization of a Multi-layer Network for Protein Function Prediction

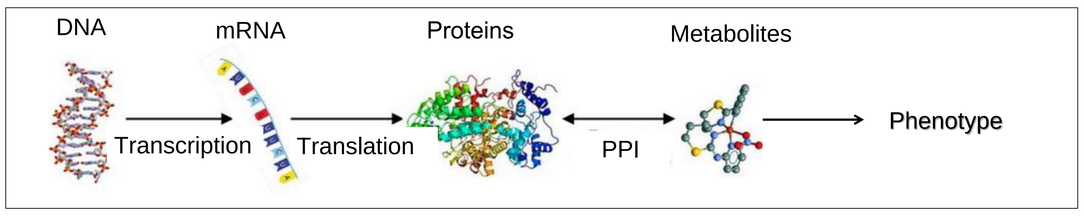
Surabhi Jagtap[1,2], Abdulkadir Çelikkanat[2], Aurélie Pirayre[1], Frederique Bidard[1], Laurent Duval[1] , Fragkiskos D. Malliaros[2]

[1]IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France

[2]Université Paris-Saclay, CentraleSupélec, Inria, Centre for Visual Computing, 91190 Gif-Sur-Yvette, France

INTRODUCTION

- Interacting entities in the organism's cellular system: **Genes**, **Proteins**, **Metabolites**, etc.
- **Omics technologies**: Enable us to study the relationships of these entities.

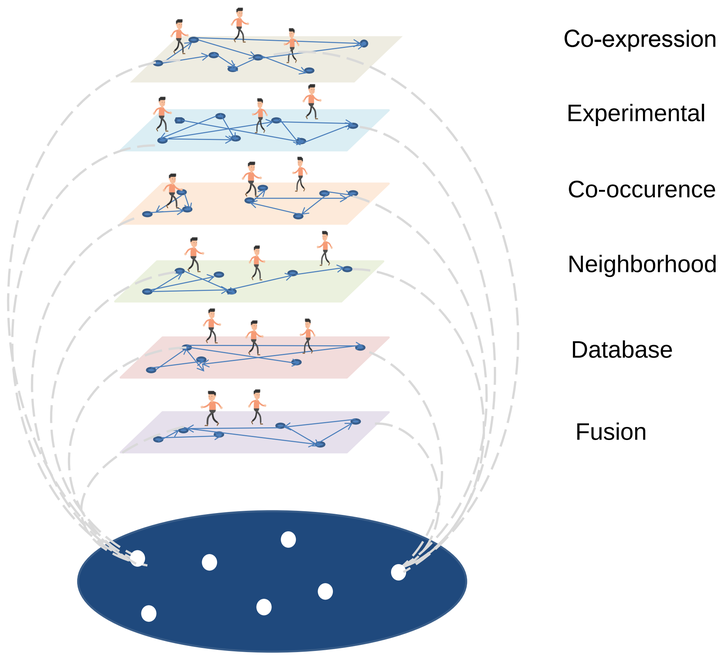


- **Biological networks**: Represent the biological relations (**edges**) between the interacting entities (**nodes**).
- Independent analysis of such networks is limited to correlations, mostly reflecting reactive processes rather than causative ones.
- **Integration** of biological networks: Can elucidate potential causative changes that can lead to a protein function prediction.
- **Graph Representation Learning (GRL)** algorithms: allow us to encode the graph structure into compact embedding vectors.

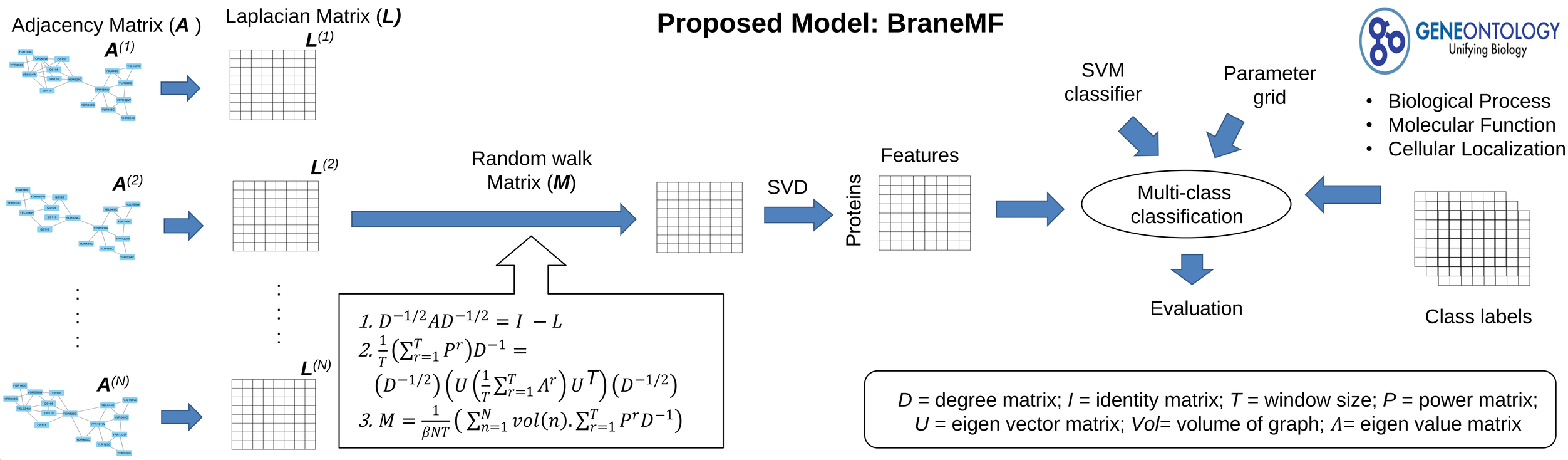
OBJECTIVE

To embed proteins of biological networks in a lower dimension latent space focusing on the application of protein function prediction

Graph Representation Learning (GRL)

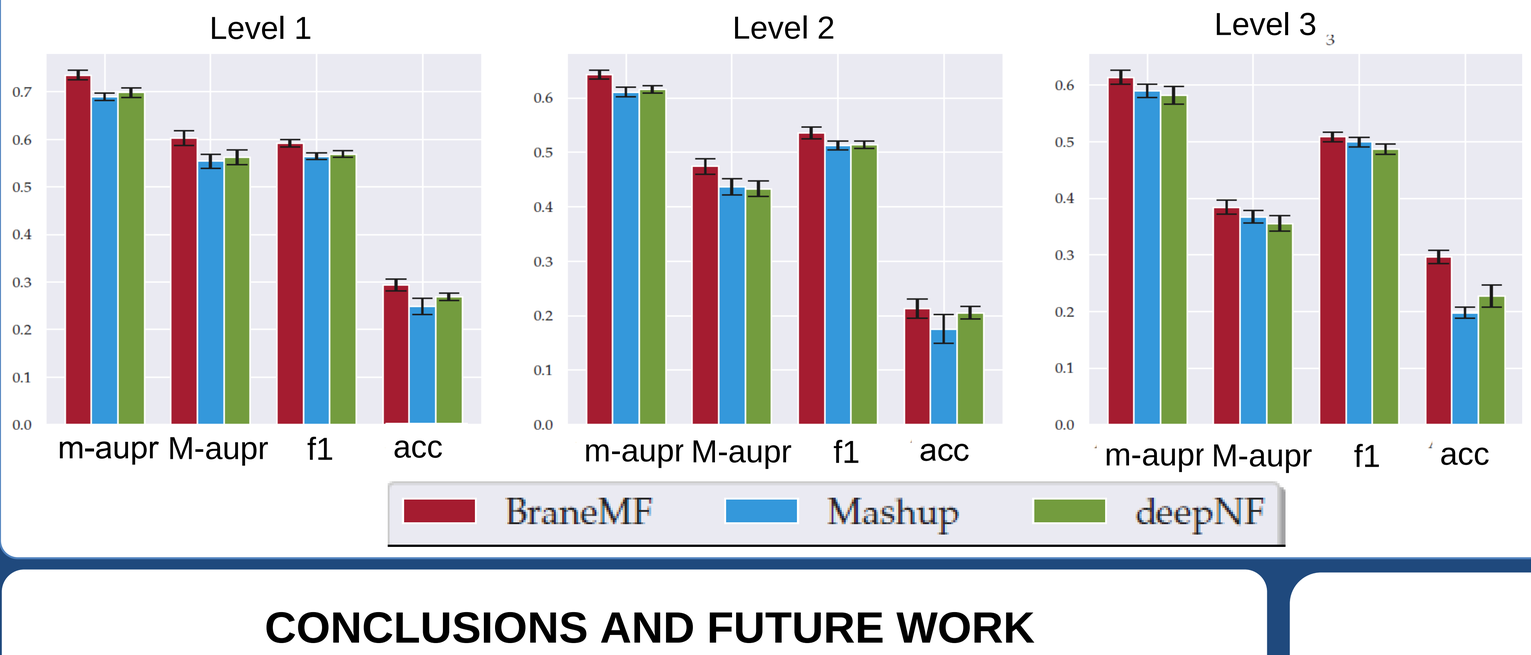


Proposed Model: BraneMF



D = degree matrix; *I* = identity matrix; *T* = window size; *P* = power matrix; *U* = eigen vector matrix; *Vol*= volume of graph; Λ= eigen value matrix

RESULTS



- Cross Validation (CV) performance of BraneMF on Gene Ontology (GO) prediction in yeast STRING networks.
- CV is compared with the performance of the state-of-the-art integration methods, Mashup and deepNF.
- Performance is measured by the area under the precision-recall curve, summarized over all GO terms both under the micro-averaging (m-aupr) and macro-averaging (M-aupr), F1 score (f1) and accuracy (acc).
- Performance shown separately for MIPS yeast annotations for Level 1, Level 2 and Level 3.
- The error bars are computed based on 10 CV trials.
- BraneMF outperforms the baseline methods in all three levels.
- The best results we can achieve is the gain of 15% in the accuracy than Mashup and 7% in deepNF.

CONCLUSIONS AND FUTURE WORK

- We have presented a graph embedding based network integration method for protein function prediction.
- We construct a compact low-dimensional protein feature representation from the selected networks. Also, we have performed an empirical analysis of the proposed approach, comparing it to state-of-art baseline methods.
- Our method is not limited to only protein function prediction tasks. There are several directions for future work with focus on extending the integration to include other data types such as protein sequences, protein structures, epigenetic marks, etc.

KEY REFERENCES

- B. Perozzi, R-Al R fou, and S. Skiena, "Deepwalk: Online learning of social representations," in Proc. 20thACM SIGKDD Int. Conf. on knowledge discovery and data mining, 2014, pp. 701–710.
- A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proc. 22nd ACM SIG KDD Int. Conf. on knowledge discovery and data mining, 2016, pp. 855–864.
- J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying DeepWalk, LINE,PTE, and node2vec," in Proceedings of the eleventh ACM international conference onweb search and data mining, 2018, pp. 459–467.
- P. Mercado, A. Gautier, F. Tudisco, and M. Hein, "The power mean Laplacian for multilayer graph clustering," in Proceedings of the Twenty First International Conference on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, A. Storkey and F. Perez Cruz, Eds., vol. 84. Playa Blanca, Lanzarote, Canary Islands: PMLR, 09–11 Apr 2018, pp. 1828–1838.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, "scikitlearn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011.
- D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. HuertaCepas, M. Simonovic, N. T. Doncheva,J. H. Morris, P. Borket al., "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome wide experimental datasets,"Nucleic acids research, vol. 47,no. D1, pp. D607–D613, 2019.
- H. Cho, B. Berger, and J. Peng, "Compact integration of multi network topology for functional analysis of genes,"Cell systems, vol. 3, no. 6, pp. 540–548, 2016.
- V. Gligorijević, M. Barot, and R. Bonneau, "deepNF: deep network fusion for protein function prediction, "Bioinformatics, vol. 34, no. 22, pp. 3873–3881, 2018.

Code: https://github.com/Surabhivj/BraneMF

Poster session: Representation Learning in Biology (RLB)

ISMB/ECCB July 2021