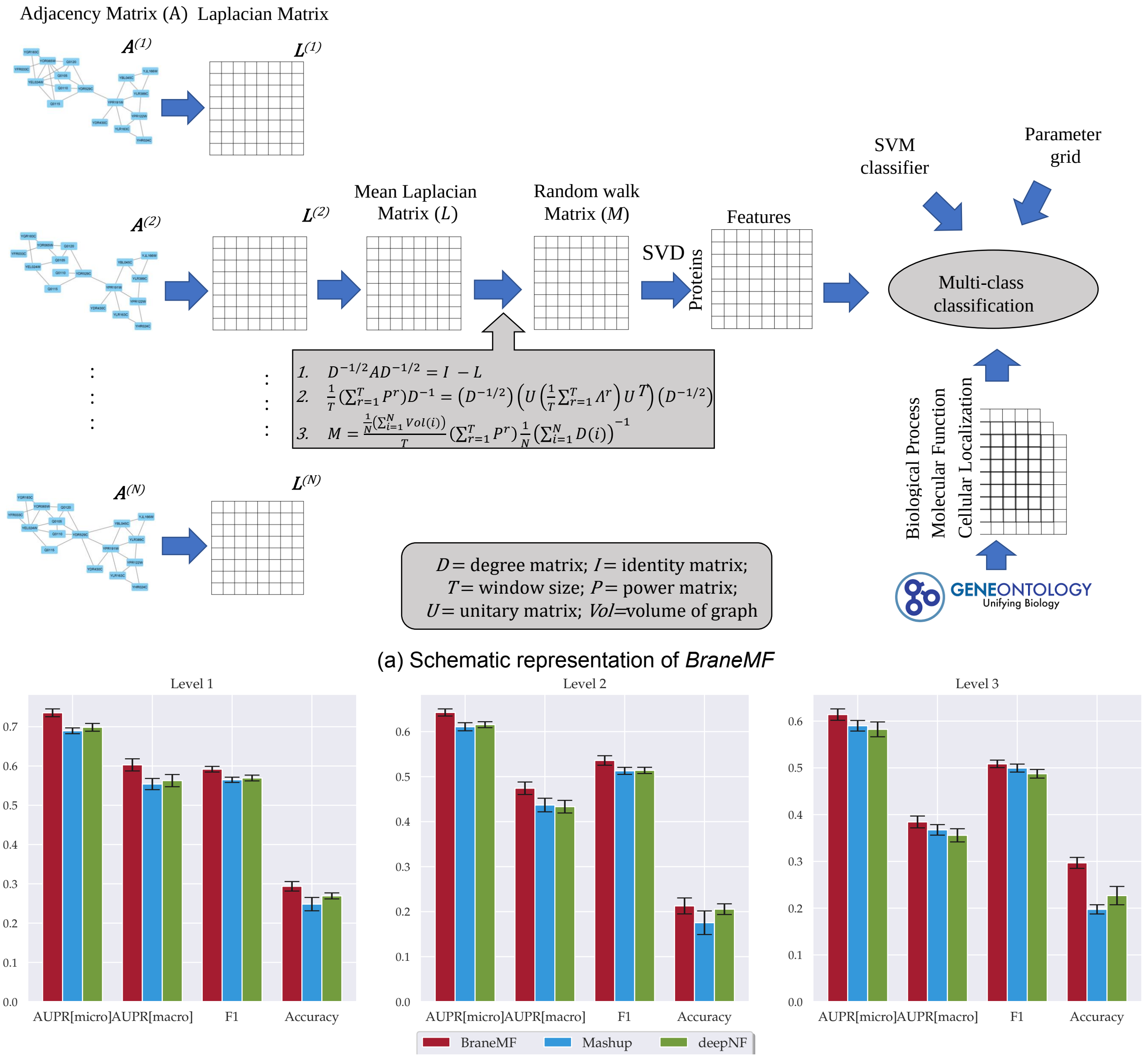
Random Walk-based Matrix Factorization of a Multi-layer Network for Protein Function Prediction

Cellular systems of organisms are composed of multiple interacting entities that control cellular processes at multiple levels by tightly regulated molecular networks. In recent years, the advent of high-throughput experimental methods has resulted in the increase of large-scale molecular and functional interaction networks [factorize networks after] such as gene co-expression networks, protein–protein interaction networks (PPI), genetic interaction networks, and metabolic networks. These networks are rich source [s] of information that could be used to infer the functional annotations of genes or proteins. Extracting relevant biological information from their topology is essential in understanding the functioning of the cell and its building blocks (proteins). Hence, to leverage protein function prediction, it is necessary to develop methods to integrate these different data sources. Although there is a plethora of graph representation learning approaches based on random walks [1; 2], matrix factorization [3], and neural networks [4], they have mostly been introduced for single-layer networks [5]. However, the heterogeneous nature of biological networks, as well as their inherent complex structure, make the development of such methods challenging. Therefore, it is necessary to obtain an informative representation of the proteins and their proximity that is not fully captured by features that are extracted directly from single input networks. Here, we focus on protein function prediction by learning node features from multiple networks.

In this work, we introduce *BraneMF*, a network integration method with the concept of matrix factorization, that generalizes random walk-based models. Network embedding as matrix factorization is a well-established method for single layer graphs to approximate the spectrum of a random walk transition matrix via the spectrum of normalized graph Laplacian [3]. We propose to blend the information of a multilayer graph Laplacian by taking certain matrix power means of Laplacian matrices for each layer [6]. Then, the normalized mean Laplacian is factorized using Singular Value Decomposition (SVD) and network embeddings are obtained by using its top-*d* singular values. We model the problem of protein function prediction as a multi-label classification problem. Further, we use the learned features to train a Support Vector Machine (SVM) classifier to predict probability scores for each protein. To evaluate the performance of the SVM on the features we adopt a 5-fold cross validation strategy. We split all annotated proteins into a training set (80%) and a test set (20%). We then train SVM on the training set and predict function of the test proteins. We use the standard radial basis kernel (RBF) for SVM and perform a nested 5-fold cross validation within the training set to select the optimal hyperparameters and the weight regularization parameter via grid search [7]. The schematic representation of the methodology is given in Figure 1a.

We test our framework with *Saccharomyces cerevisiae* (yeast), a well-studied micro-organism. We used STRING PPI networks [8] from six different data sources such as co-expression, co-occurence, experimental, database, fusion and neighborhood. We compare the performance of *BraneMF* against the state-of-the-art integration methods for protein function prediction *Mashup* [9] and *deepNF* [10]. *Mashup* is a network integration framework based on matrix factorization that builds compact low-dimensional vector representation of proteins. *deepNF* is a network fusion method based on multimodal deep autoencoders. Both approaches consider a set of input networks and for each network they construct vector representations of proteins. We use cross validation to evaluate the classification performance of our model (Figure 1b). From our preliminary results, it is observed that *BraneMF* could capture relevant protein features more accurately from complex heterogeneous PPI networks as compared to the state-of-art methods. The best results we can achieve is the gain of 15% in the accuracy than *Mashup* and 7% in *deepNF*. The unsupervised learned features are independent of downstream machine learning tasks, the learned embeddings could be leveraged across a wide variety of multi-network integration tasks such as Gene Regulatory Network (GRN) inference and Transcription Factor (TF) target identification.


Adjacency Matrix (A)
Laplacian Matrix
$A^{(1)}$
$L^{(1)}$
$A^{(2)}$
$L^{(2)}$
$A^{(N)}$
$L^{(N)}$
Mean Laplacian Matrix (L)
Random walk Matrix (M)
SVD
Proteins
Features
SVM classifier
Parameter grid
Multi-class classification
Biological Process
Molecular Function
Cellular Localization
GENEONTOLOGY
Unifying Biology
1. $D^{-1/2}AD^{-1/2} = I - L$
2. $\frac{1}{T}(\sum_{r=1}^{T} P^r)D^{-1} = \left(D^{-1/2}\right)\left(U\left(\frac{1}{T}\sum_{r=1}^{T} \Lambda^r\right)U^T\right)\left(D^{-1/2}\right)$
3. $M = \frac{\frac{1}{N}(\sum_{i=1}^{N} Vol(i))}{T}(\sum_{r=1}^{T} P^r)\frac{1}{N}\left(\sum_{i=1}^{N} D(i)\right)^{-1}$
D = degree matrix; I = identity matrix;
T = window size; P = power matrix;
U = unitary matrix; Vol=volume of graph
Level 1
Level 2
Level 3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
AUPR[micro]
AUPR[macro]
F1
Accuracy
BraneMF
Mashup
deepNF

(a) Schematic representation of *BraneMF*

(b) Protein function prediction

Figure 1: **(a). Schematic representation of *BraneMF*:** In the first step, the input networks are converted into their Laplacian matrix. The mean of this Laplacian is normalized and approximated to random walk matrix. The random walk matrix is factorized using SVD. Low-dimensional features are then extracted by using its top-d singular values. These protein features are used to predict the functional annotations of proteins. **(b). Preliminary results:** Cross-validation performance of *BraneMF* on function prediction in *yeast* STRING networks. It is compared with function prediction 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 functional annotation terms both under the micro-averaging (AUPR [micro]) and macro-averaging (AUPR [macro]), F1 score and accuracy. Performance of the methods is shown separately for MIPS *yeast* annotations for Level 1 (left), Level 2 (middle) and Level 3 (right). The error bars are computed based on 10 cross-validation trials. The results show that *BraneMF* outperforms *deepNF* and *Mashup*.

References

[1] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proc. 20th ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, 2014, pp. 701–710.

[2] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *Proc. 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining*, 2016, pp. 855–864.

[3] 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 on web search and data mining*, 2018, pp. 459–467.

[4] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” *IEEE Data Eng. Bull.*, vol. 40, no. 3, pp. 52–74, 2017.

[5] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S. M. Lin, W. Zhang, and P. Zhang, “Graph embedding on biomedical networks: methods, applications and evaluations,” *Bioinformatics*, vol. 36, pp. 1241–1251, Feb. 2020.

[6] 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.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, “scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, 2011.

[8] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork *et 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.

[9] 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.

[10] V. Gligorijević, M. Barot, and R. Bonneau, “deepNF: deep network fusion for protein function prediction,” *Bioinformatics*, vol. 34, no. 22, pp. 3873–3881, 2018.