Learning topology-preserving embedding for gene interaction networks

Kishan K C (kk3671@rit.edu)[1], Rui Li[1], Feng Cui[2], Anne R. Haake[1]

[1]Department of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY 14623,USA

[2]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY 14623,USA

ECCB 2018

Introduction

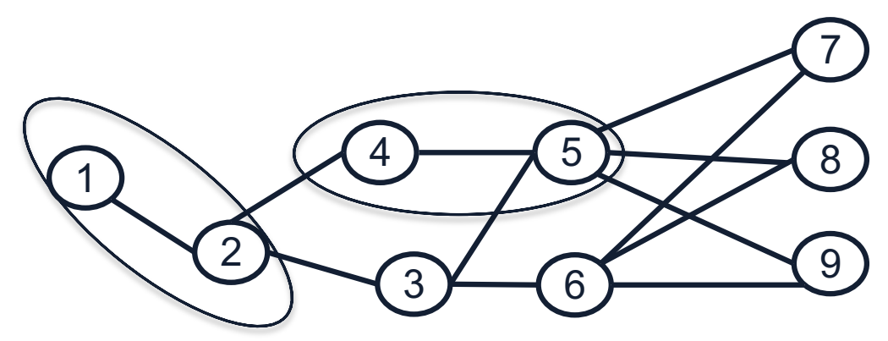
- Understanding **functional aspects of genes or proteins** is crucial to provide insights into underlying biological phenomenon for different health and disease conditions.
- **Often intractable** through biological experiments.
- Propose a deep neural network architecture to learn lower dimensional representation [1] for each gene, by preserving the topological properties of gene interaction network.
- By preserving network topology, this approach place genes with similar topological patterns closer to each other in embedding space.
- Encoding genes to lower dimensional representation will assist tasks like gene function prediction, genetic interaction prediction and gene ontology reconstruction.
- We show that our model learns comprehensive representation of network topology of gene interaction networks that improves the performance in genetic interaction prediction for yeast and ecoli datasets.

Background

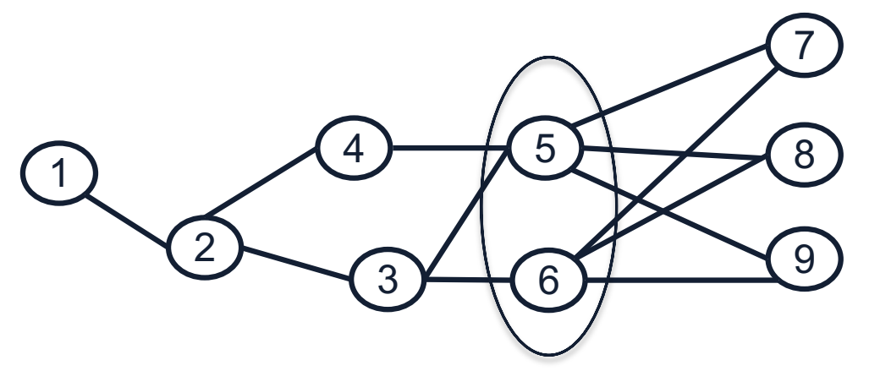
Gene Network can be defined as graph **G = (V, E)** where
V = {V_1, V_2,….,V_M} denotes the genes or proteins,
E = {e_{ij}} is the neighborhood relationship between genes, and

Structural Proximity : the proximity of nodes in network structure.

Direct Proximity



Indirect Proximity



Given a gene network denoted as G = (V, E), gene network embedding aims to learn a function **f** that maps topological properties of gene v_i to d-dimensional vector y where d << |V|. The objective of function f is to learn low dimensional vector y_i and y_j for gene v_i and v_j such that the similarity between them explicitly preserves the topological similarity.

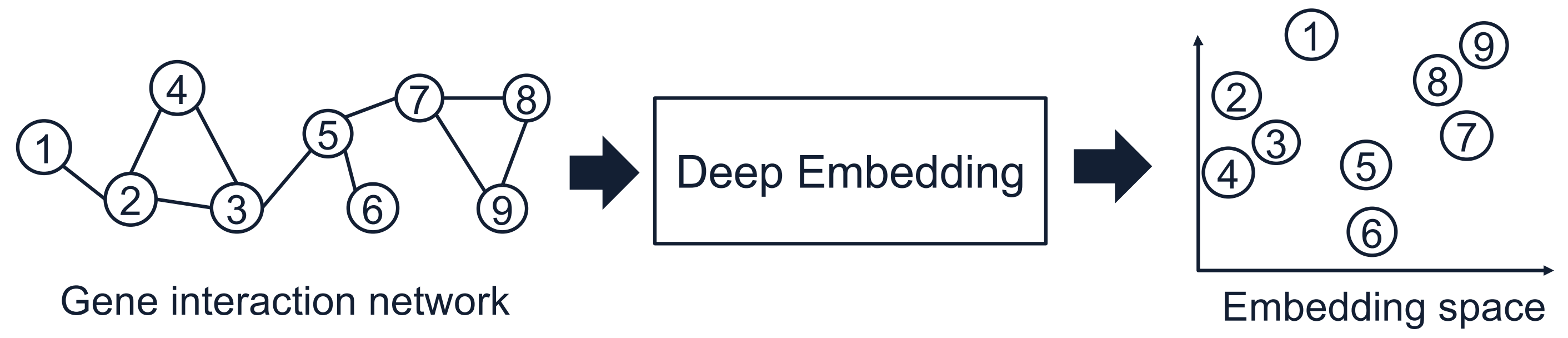
Datasets

- Interaction network data from BioGRID database [2].

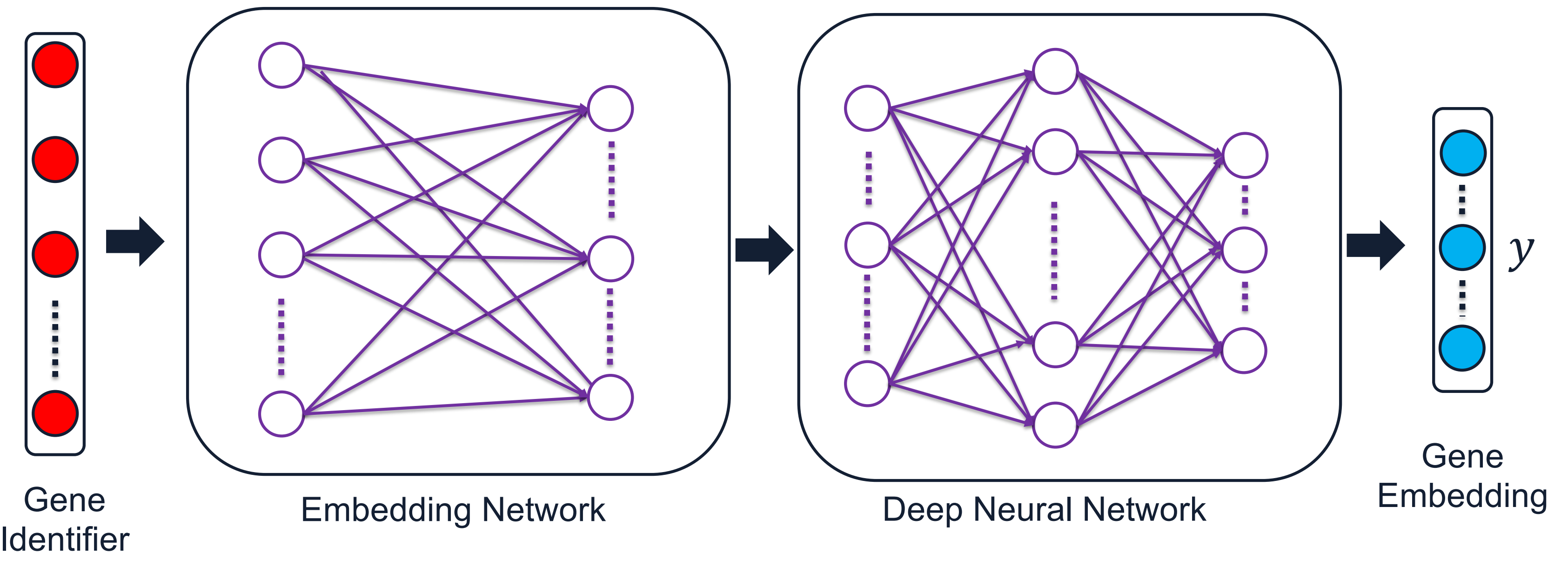
	Interaction Network Data	
Organism	**# (Genes)**	**# (Interactions)**
Yeast	5,950	544,652
Ecoli	4,511	148,340

Topology-preserving Embedding Model

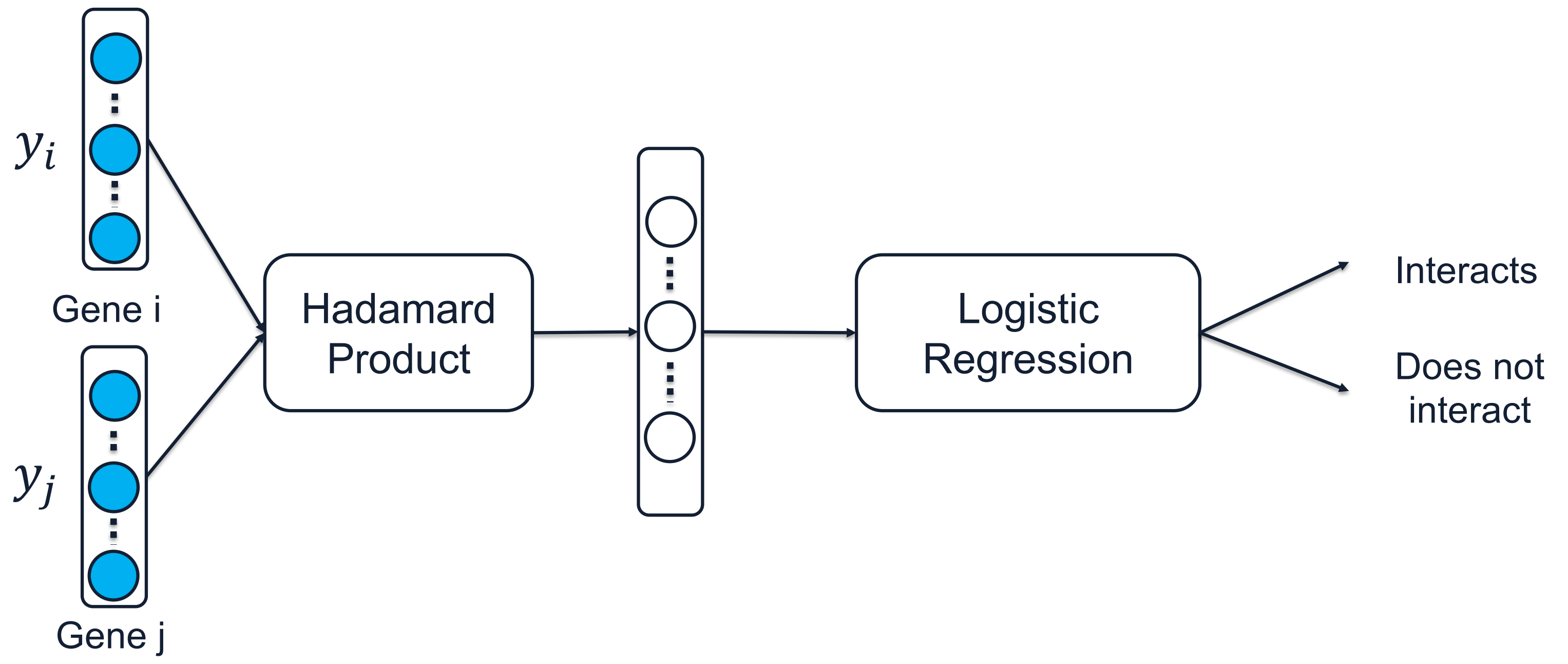
- Topological Information: Gene Interaction Network
- Learns low dimensional representation for each gene, preserving topological proximity between genes.



- Deep Neural Network to model topological proximity for gene interaction network.



- Learning binary classifier to separate interacting gene pairs from non-interacting pairs.



Acknowledgements

This material is based upon work supported by National Science Foundation under Grant NSF-1062422.

Results: Genetic Interaction Prediction

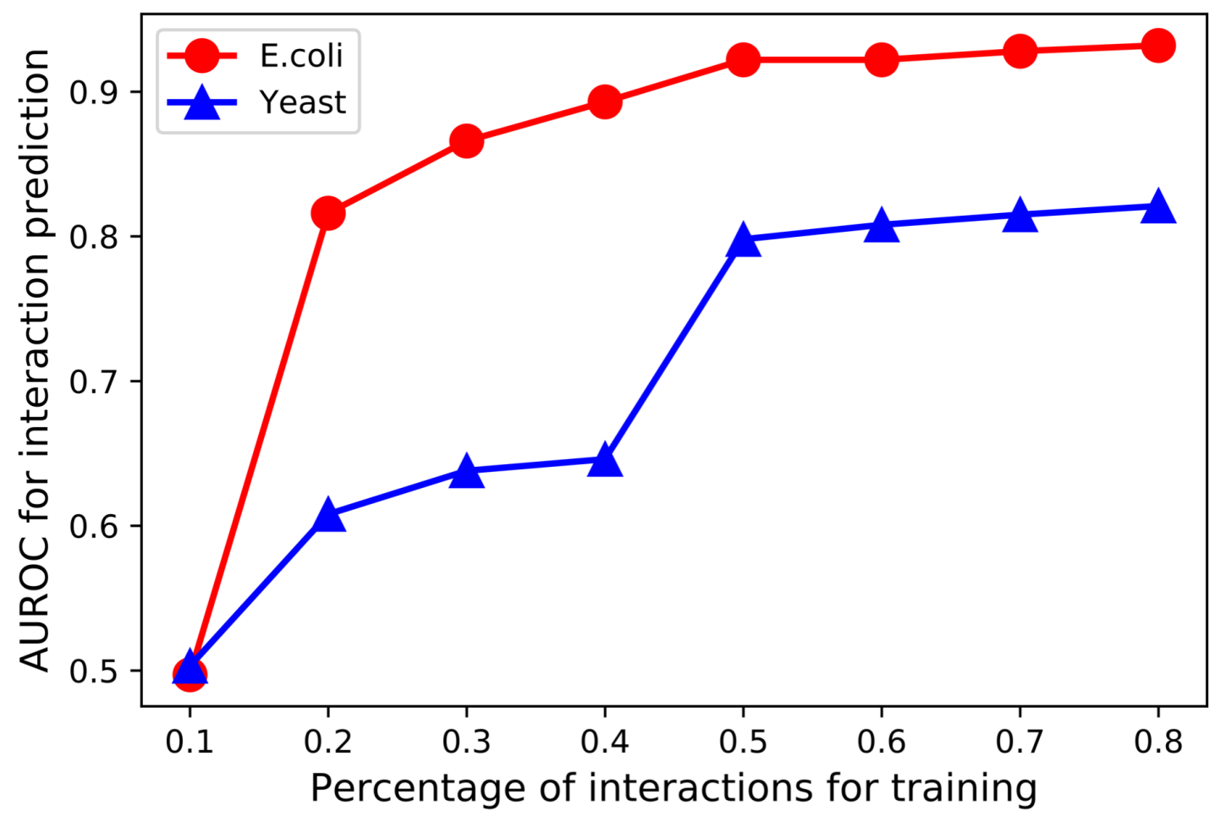
- Optimal parameters

Dataset	Learning rate	Batch size	λ	d
Yeast	0.005	128	0.8	128
Ecoli	0.005	64	1.0	128

- AUROC comparison shows that GNE outperforms other methods.

Methods	Yeast		Ecoli	
	AUROC	**AUPR**	**AUROC**	**AUPR**
Isomap [3]	0.507	0.588	0.559	0.672
LINE [4]	0.726	0.686	0.897	0.851
node2vec [5]	0.739	0.708	0.912	0.862
Our method	**0.787**	**0.784**	**0.930**	**0.931**

- Performance of our model depends on the percentage of interactions taken for training the model.



Conclusion

- Our method can learn effective representation for gene interaction networks that can be used to infer unknown gene interactions.
- Future work includes integration of other information about genes like gene expression, functional annotations, sequence similarity, functional information etc. [6] and evaluation of gene embedding for gene function prediction and gene ontology reconstruction.

References

[1] Hamilton, William L., Rex Ying, and Jure Leskovec. "Representation Learning on Graphs: Methods and Applications." *arXiv preprint arXiv:1709.05584* (2017).

[2] Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. "BioGRID: a general repository for interaction datasets." Nucleic acids research 34, no. suppl_1 (2006): D535-D539.

[3] Lei, Ying-Ke, et al. "Assessing and predicting protein interactions by combining manifold embedding with multiple information integration." BMC bioinformatics. Vol. 13. No. 7. BioMed Central, 2012.

[4] Tang, Jian, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. "Line: Large-scale information network embedding." In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067-1077. International World Wide Web Conferences Steering Committee, 2015.

[5] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855-864. ACM, 2016.

[6] Madhukar, Neel S., Olivier Elemento, and Gaurav Pandey. "Prediction of genetic interactions using machine learning and network properties." Frontiers in bioengineering and biotechnology 3 (2015): 172.