GNE: A deep learning framework for gene network inference by aggregating biological information

Kishan K C

Human-Centric Multi-Modal Modeling Lab

Golisano College of Computing and Information Sciences

Rochester Institute of Technology, New York, USA

Background

- Gene interaction network is a set of genes (nodes) connected by edges representing functional relationships among these genes.

- Interactions are important to
    - Understand pathways and regulation in model organisms
    - Understand biological functions
    - Provide Insight into complex diseases

Intractable through biological experiments

Background

- Advancement in measurement technologies => large amount of high-throughput datasets
- Topological properties of gene interaction network
- Guilt by association: allows to discover similar genes but also to infer the properties of unknown ones
- Proposed methods: Isomap (*Lei et al. 2012*), node2vec (*Grover & Leskovec 2016*), LINE (*Tang et al. 2015*)

Background

- Preserving topological information is not enough
- Some of the genes have no interaction information
- Genes with similar attributes are likely to be related

Our approach:

Integrate topological properties and additional information

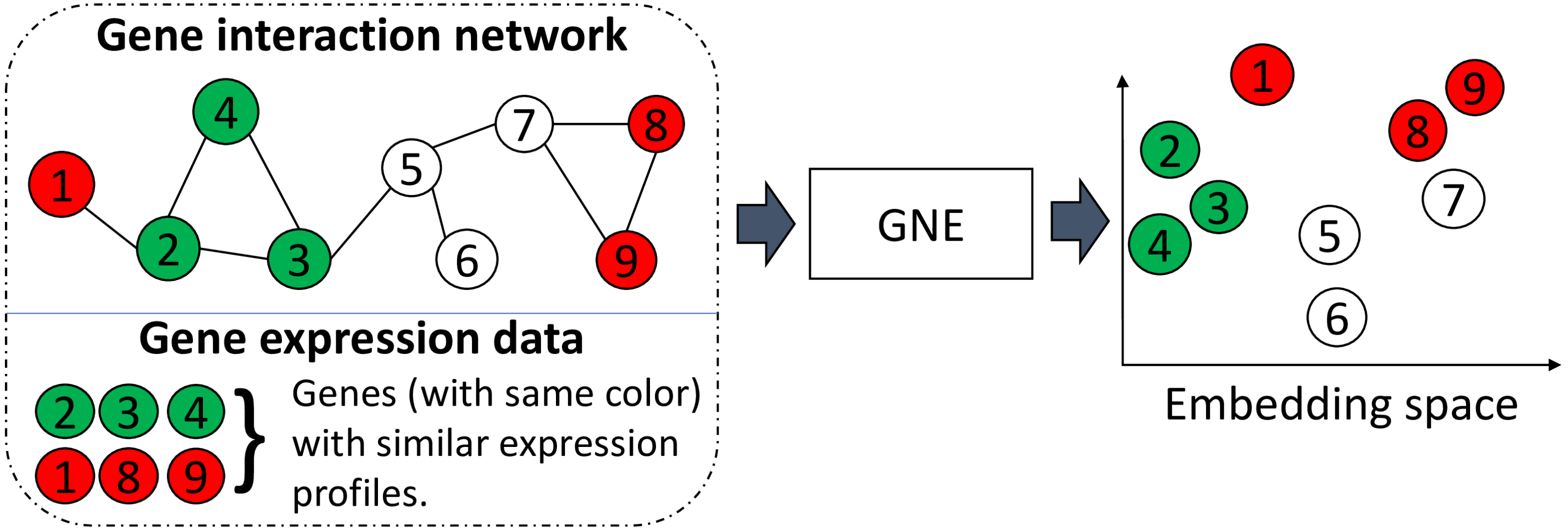
Datasets

- Interaction dataset from BioGRID database (*Stark et al. 2006*)
- Gene Expression data from the DREAM5 Network Challenge (*Marbach et al. 2012*)
- Operons dataset from the DOOR database (*Mao et al. 2008*)

Gene Network Embedding (GNE)

- A novel deep learning framework to integrate diverse biological information for GI network inference

- Incorporates gene expression data with GI network topological information



Gene interaction network

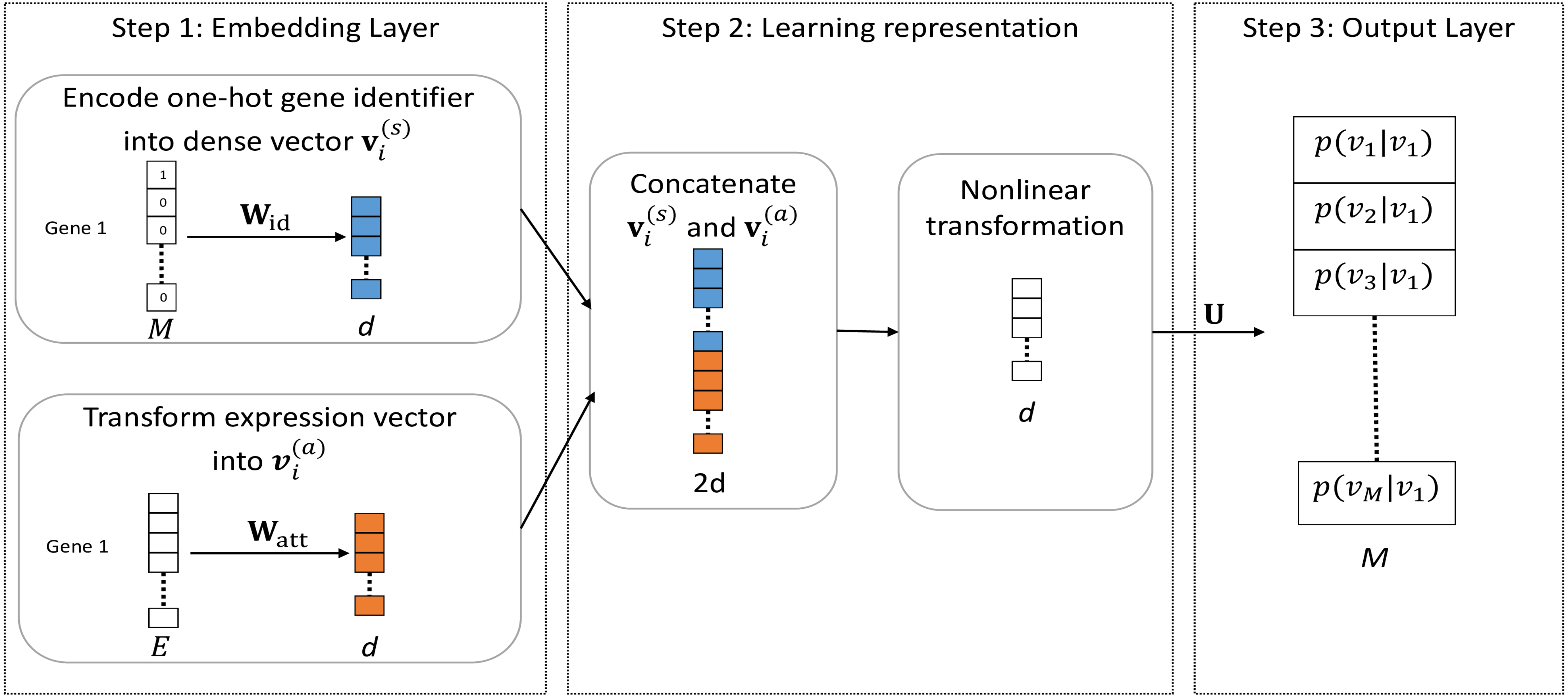
Gene expression data

Genes (with same color) with similar expression profiles.

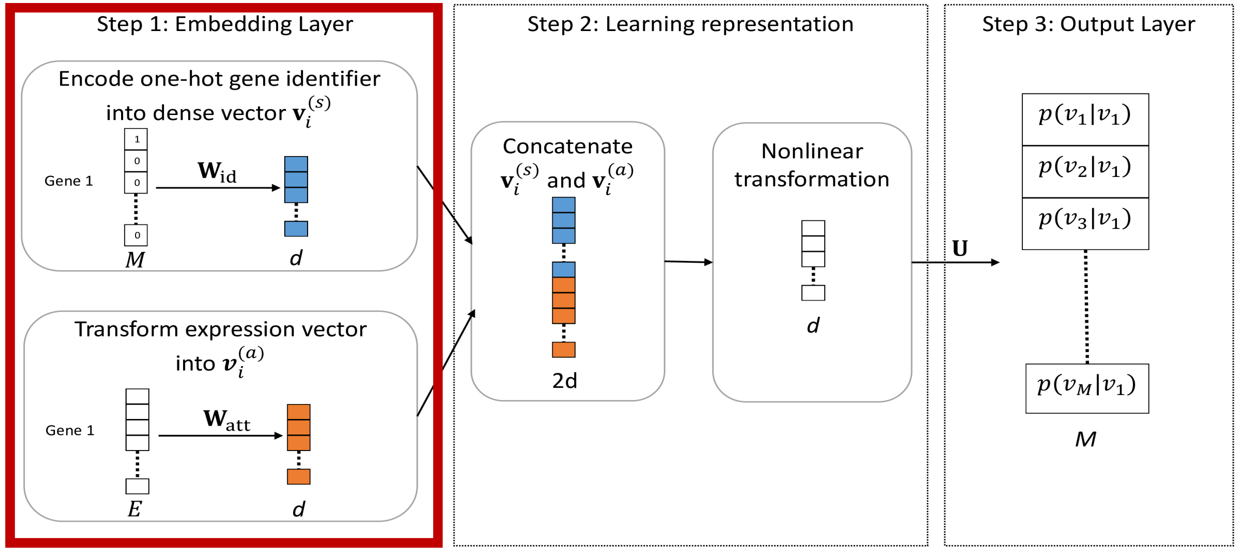
GNE

Embedding space

GNE Architecture


Step 1: Embedding Layer
Encode one-hot gene identifier into dense vector $\mathbf{v}_i^{(s)}$
Gene 1
1
0
0
0
$\mathbf{W}_{\text{id}}$
M
d
Transform expression vector into $\boldsymbol{v}_i^{(a)}$
Gene 1
$\mathbf{W}_{\text{att}}$
E
d
Step 2: Learning representation
Concatenate $\mathbf{v}_i^{(s)}$ and $\mathbf{v}_i^{(a)}$
2d
Nonlinear transformation
d
Step 3: Output Layer
$\mathbf{U}$
$p(v_1|v_1)$
$p(v_2|v_1)$
$p(v_3|v_1)$
$p(v_M|v_1)$
M

GNE: Embedding


Step 1: Embedding Layer
Encode one-hot gene identifier into dense vector $\mathbf{v}_i^{(s)}$
Gene 1
1
0
0
0
$\mathbf{W}_{\text{id}}$
M
d
Transform expression vector into $\boldsymbol{v}_i^{(a)}$
Gene 1
$\mathbf{W}_{\text{att}}$
E
d
Step 2: Learning representation
Concatenate $\mathbf{v}_i^{(s)}$ and $\mathbf{v}_i^{(a)}$
2d
Nonlinear transformation
d
Step 3: Output Layer
$\mathbf{U}$
$p(v_1|v_1)$
$p(v_2|v_1)$
$p(v_3|v_1)$
$p(v_M|v_1)$
M

GNE Network Structure Modeling

Encode one-hot encoded representation of a gene v_i via embedding lookup.

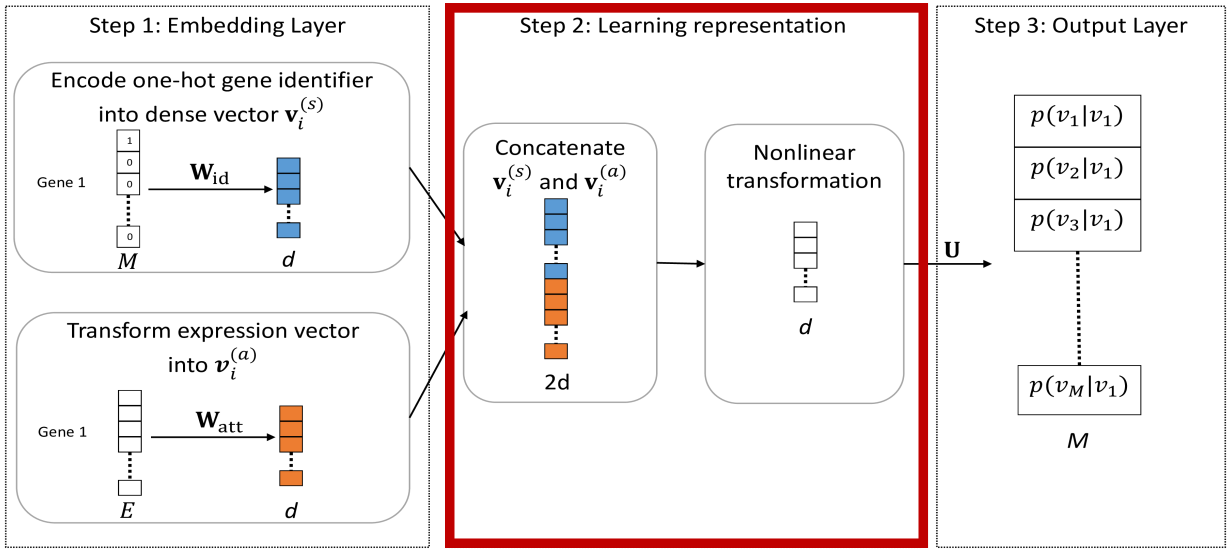
$$\mathbf{v}_i^{(s)} = \boldsymbol{W_{id}}\, v_i$$

GNE Expression Modeling

Exponential Linear unit (ELU) to model non-linearity of gene expression x_i and capture underlying patterns.

$$\mathbf{v}_i^{(a)} = \text{elu}(\ \boldsymbol{W_{att}} \cdot x_i)$$

GNE: Learning representation



Step 1: Embedding Layer
Encode one-hot gene identifier into dense vector $\mathbf{v}_i^{(s)}$
Gene 1
1
0
0
0
$\mathbf{W}_{\text{id}}$
M
d
Transform expression vector into $\boldsymbol{v}_i^{(a)}$
Gene 1
$\mathbf{W}_{\text{att}}$
E
d
Step 2: Learning representation
Concatenate $\mathbf{v}_i^{(s)}$ and $\mathbf{v}_i^{(a)}$
2d
Nonlinear transformation
d
$\mathbf{U}$
Step 3: Output Layer
$p(v_1|v_1)$
$p(v_2|v_1)$
$p(v_3|v_1)$
$p(v_M|v_1)$
M

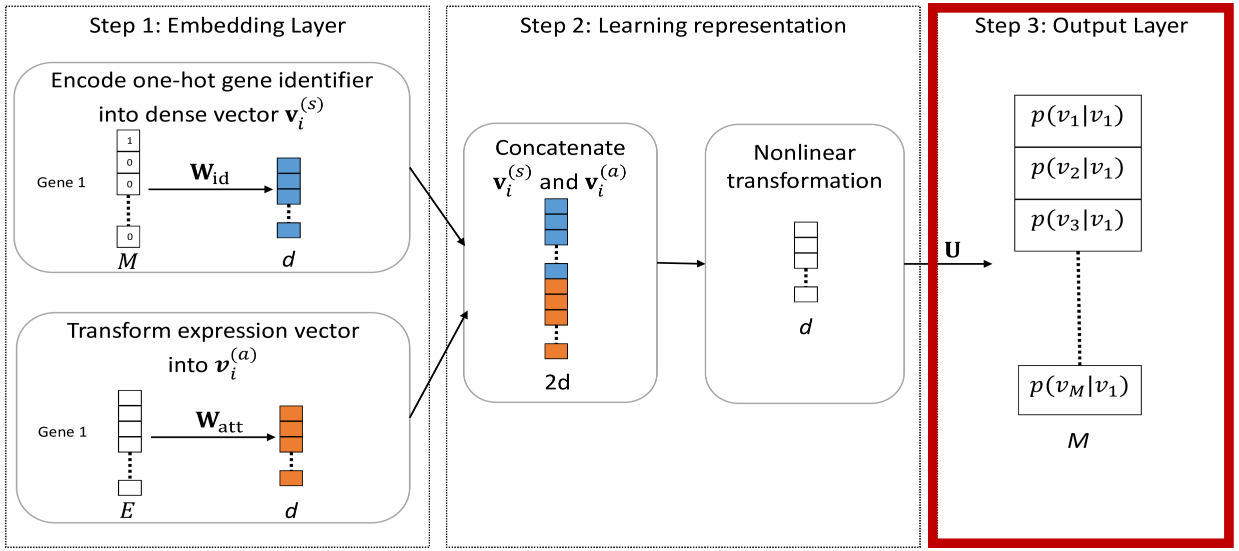
Concatenation of topological and attribute representation

$$\mathbf{v}_i = [\mathbf{v}_i^{(s)} \quad \lambda \mathbf{v}_i^{(a)}]$$

Transformation of concatenated representation via ***k***-hidden layers with hyperbolic tangent activation.

$$\mathbf{h}_i^{(k)} = \delta_k(\boldsymbol{W_k}\, \mathbf{h}_i^{(k-1)} + b^{(k)})$$

GNE: Predicting interaction probabilities



Step 1: Embedding Layer
Encode one-hot gene identifier into dense vector $\mathbf{v}_i^{(s)}$
Gene 1
1
0
0
0
$\mathbf{W}_{\text{id}}$
M
d
Transform expression vector into $\boldsymbol{v}_i^{(a)}$
Gene 1
$\mathbf{W}_{\text{att}}$
E
d
Step 2: Learning representation
Concatenate $\mathbf{v}_i^{(s)}$ and $\mathbf{v}_i^{(a)}$
2d
Nonlinear transformation
d
Step 3: Output Layer
$\mathbf{U}$
$p(v_1|v_1)$
$p(v_2|v_1)$
$p(v_3|v_1)$
$p(v_M|v_1)$
M

Last layer outputs the probability vector which contains conditional probability of all other genes to gene v_i

$$\mathbf{o}_i = [p(v_1|v_i), p(v_2|v_i), \ldots, p(v_M|v_i)]$$

where

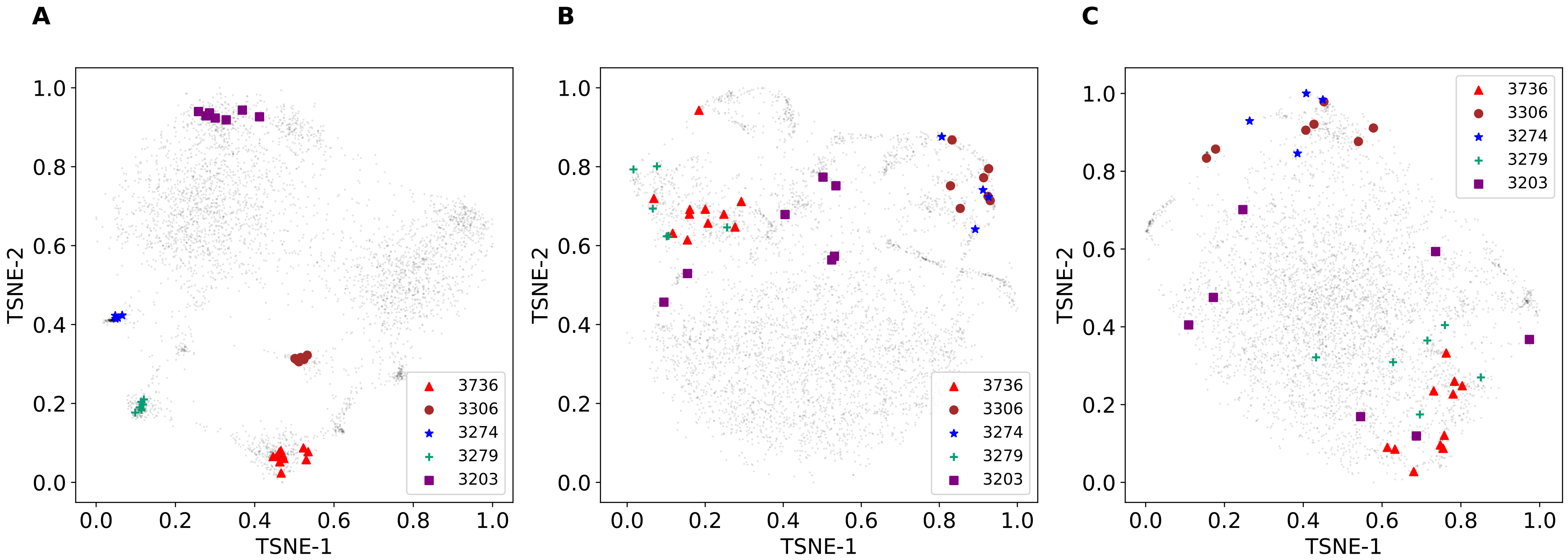
$$p(v_j|v_i) = \frac{\exp(\widetilde{\mathbf{v}}_j \cdot \mathbf{h}_i^{(k)})}{\sum_{j'=1}^{M} \exp(\widetilde{\mathbf{v}}_{j'} \cdot \mathbf{h}_i^{(k)})}$$

Optimization:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \left[\sum_{i=1}^{M} \sum_{v_j \,\in\, N_i} \log \frac{\exp(\widetilde{\mathbf{v}}_j \cdot \mathbf{h}_i^{(k)})}{\sum_{j'=1}^{M} \exp(\widetilde{\mathbf{v}}_{j'} \cdot \mathbf{h}_i^{(k)})} \right]$$

Visualizing the embeddings

- Visualize embeddings on 2D space using t-SNE package
- Operons: genes that interact with each other and are co-regulated.
  - Colored the points in 2D space with operons

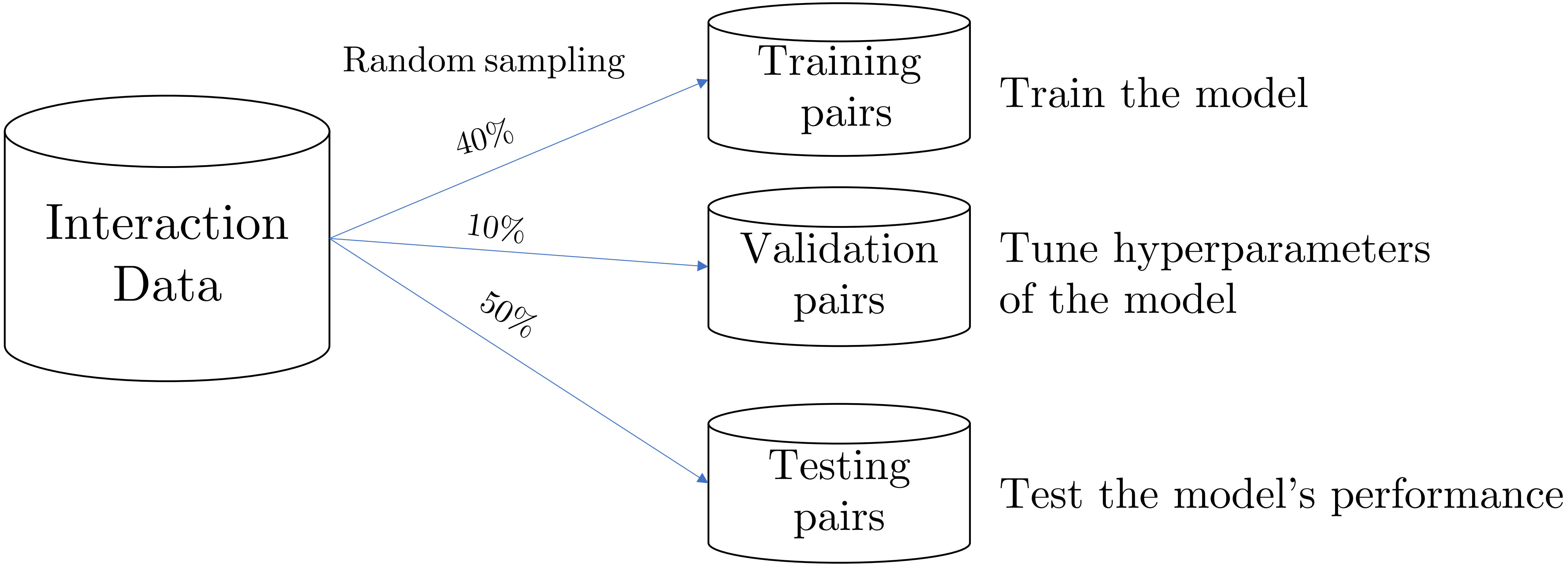


GNE LINE node2vec

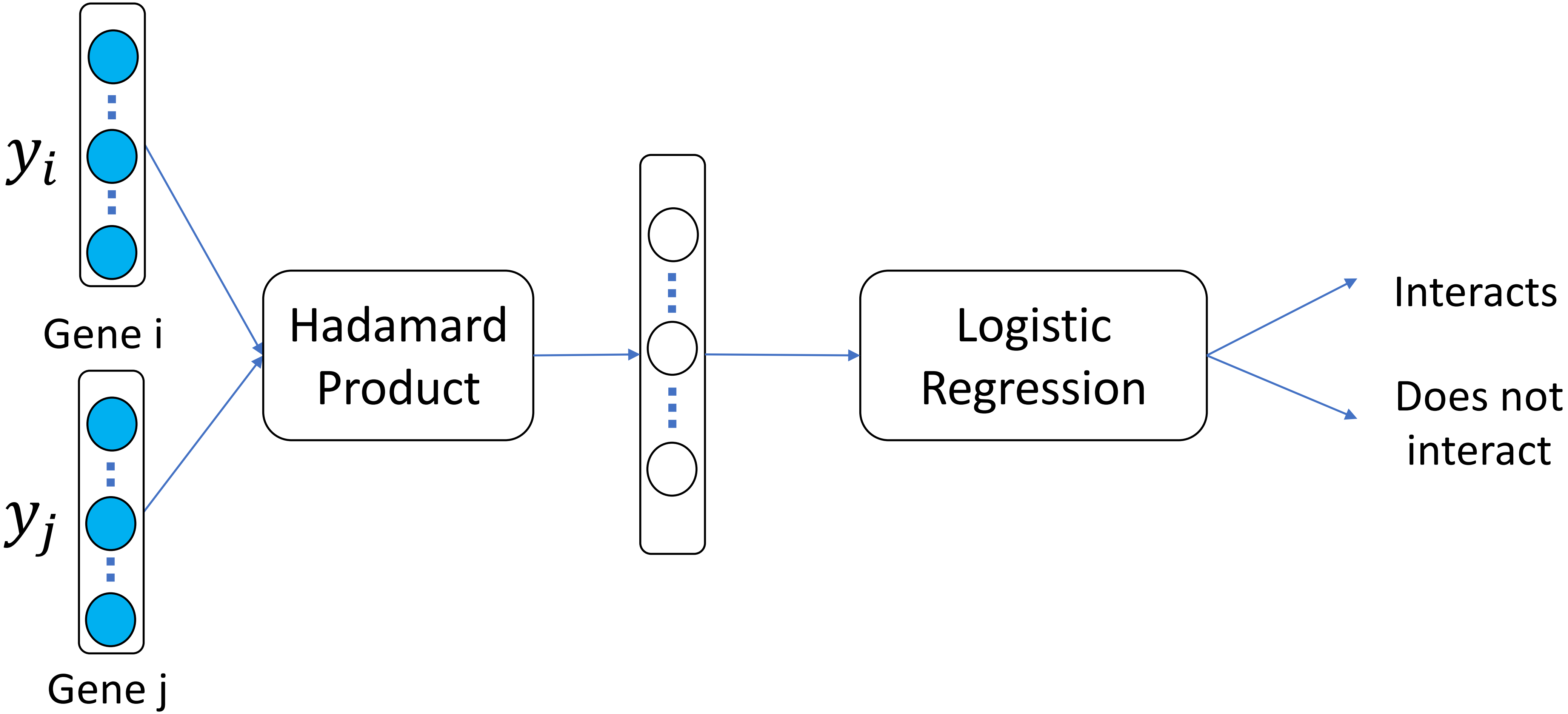
- Significant test to see if genes within same operons are likely to have similar representation

Experimental setup

- Splitting data



Training Classifier



- Random selection of negative samples

Results

- Performance of GNE in predicting missing interactions

Methods	Yeast		E. coli	
	AUROC	AUPR	AUROC	AUPR
Correlation	0.582	0.579	0.537	0.557
Isomap	0.507	0.588	0.559	0.672
LINE	0.726	0.686	0.897	0.851
node2vec	0.739	0.708	0.912	0.862
Isomap+	0.653	0.652	0.644	0.649
LINE+	0.745	0.713	0.899	0.856
node2vec+	0.751	0.716	0.871	0.826
GNE (topology only)	0.787	0.784	0.930	0.931
GNE	**0.825**	**0.821**	**0.940**	**0.939**

Temporal holdout validation

- Two versions of interaction dataset: 2017 and 2018 version
  - 2018 version has 12,835 new interactions for yeast and 11,185 new interactions for E. coli
- Randomly selected 50% of interactions from 2017 version as training data to predict new interactions in 2018 version

Methods	Yeast		E. coli	
	AUROC	AUPR	AUROC	AUPR
LINE	0.620	0.611	0.569	0.598
node2vec	0.640	0.609	0.587	0.599
GNE	**0.710**	**0.683**	**0.653**	**0.658**

GNE's predictions

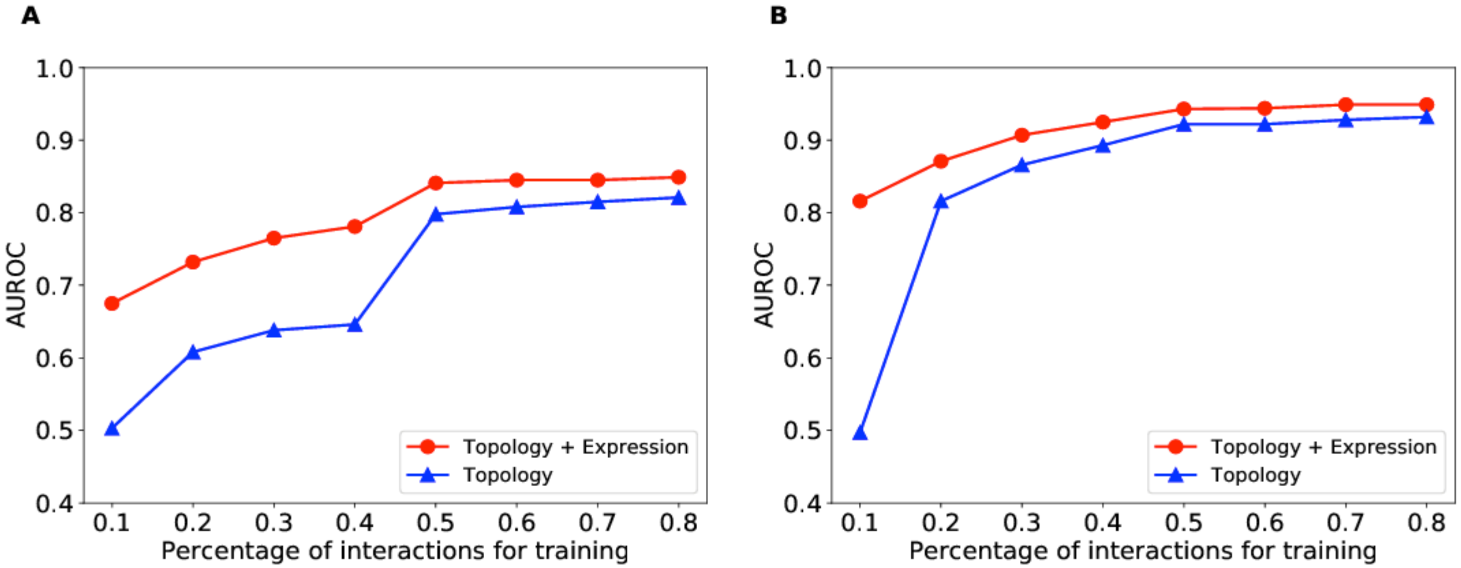
- Trained GNE with and without expression data
- Improved predictions with expression data

Dataset	Probability		Gene i	Gene j	Experimental Evidence code
	Topology	Topology + Expression			
Yeast	0.287	0.677	TFC8	DHH1	Affinity Capture-RNA[1]
	0.394	0.730	SYH1	DHH1	Affinity Capture-RNA[1]
	0.413	0.746	CPR7	DHH1	Affinity Capture-RNA[1]
E. coli	0.014	0.944	ATPB	RFBC	Affinity Capture-MS[2]
	0.012	0.941	NARQ	CYDB	Affinity Capture-MS[2]
	0.013	0.937	PCNB	PAND	Affinity Capture-MS[2]

[1] *Miller, J. E. et al. 2018* [2] *Babu, M. et al. 2018*

Impact of network sparsity

- Hold out 10% interactions as test dataset
- Change the sparsity of training data by randomly removing a portion of remaining interactions
- Evaluation with and without expression data

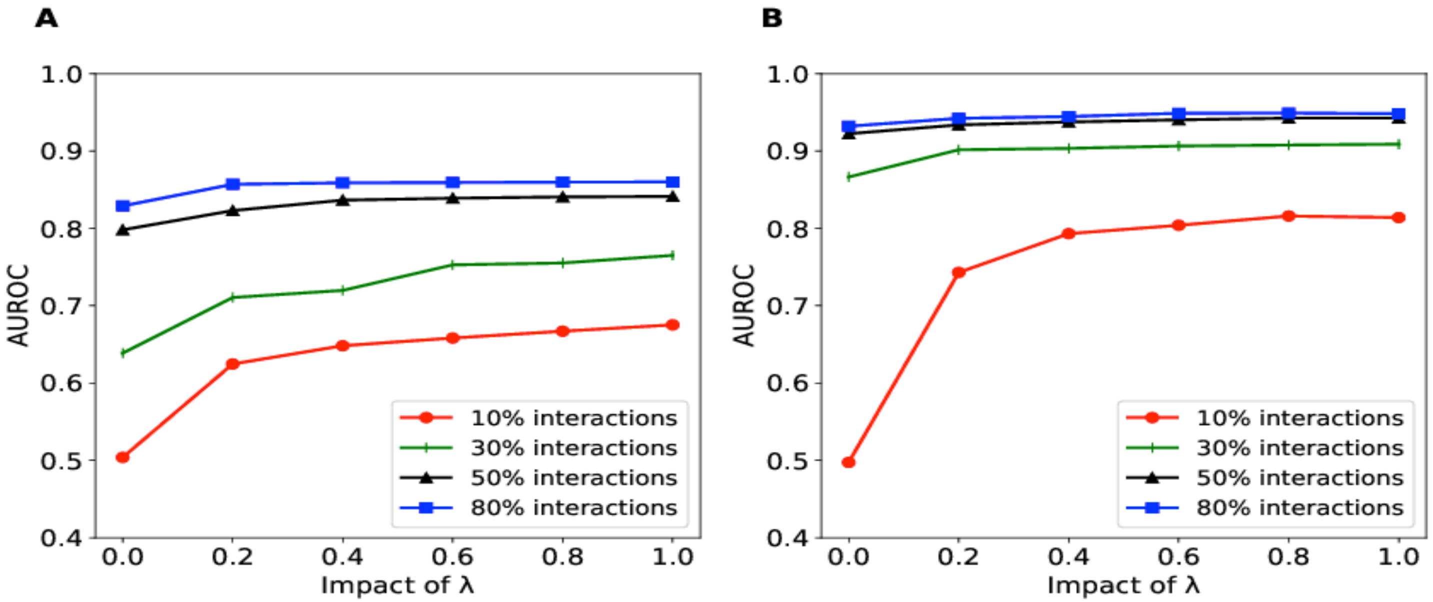


Yeast

E. coli

Impact of λ

- Evaluation of parameter λ to see the impact on model's performance
- Values of λ used in experiment: [0, 0.2, 0.4, 0.6, 0.8, 1, 10, 100, 1000]



Yeast

E. coli

Conclusion

- GNE models the complex statistical relationships between gene interaction network and expression data.

- GNE extracts features that are more informative for interaction prediction.

- GNE allows the addition of different types of attributes.

Acknowledgements

Co-authors

- Rui Li
- Feng Cui
- Qi Yu
- Anne R. Haake

Funding

NSF

NIH National Institutes of Health
Turning Discovery Into Health

Thanks