Enzymatic link prediction for biochemical route synthesis via learning graph representations of biochemical networks

Julie Jiang[1], Li-Ping Liu[1], and Soha Hassoun[1,2]
[1]Department of Computer Science, Tufts University, Medford, MA
[2]Department of Chemical and Biological Engineering, Tufts University, Medford, MA

Motivation

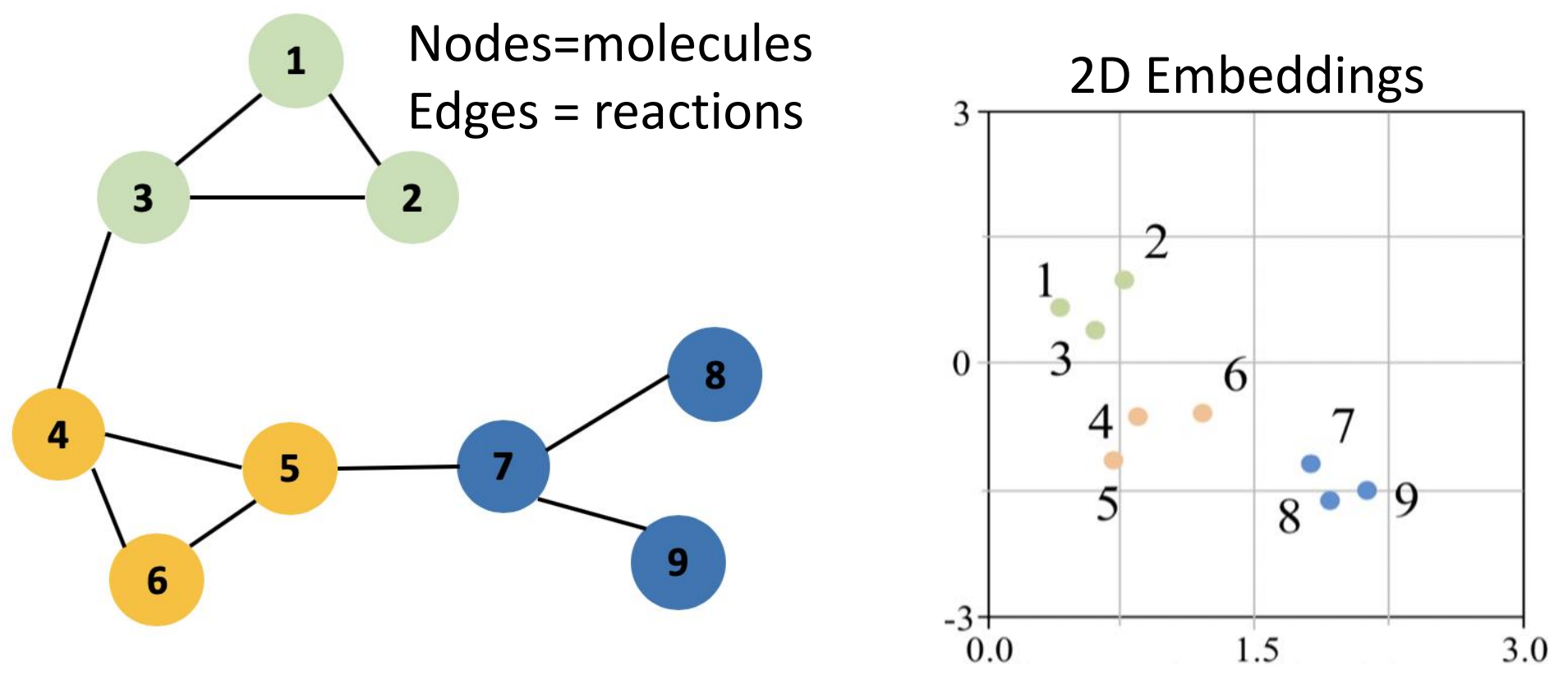
Problem: No complete characterization of enzymatic reactions
The curation of enzyme functions and the reactions they catalyze remains elusive, hindering biological engineering and discovery.

Goal: Predict enzymatic transformations

- Enhance biological discovery of undocumented enzymatic reactions
- Plan synthesis routes using previously undocumented enzymatic transformations

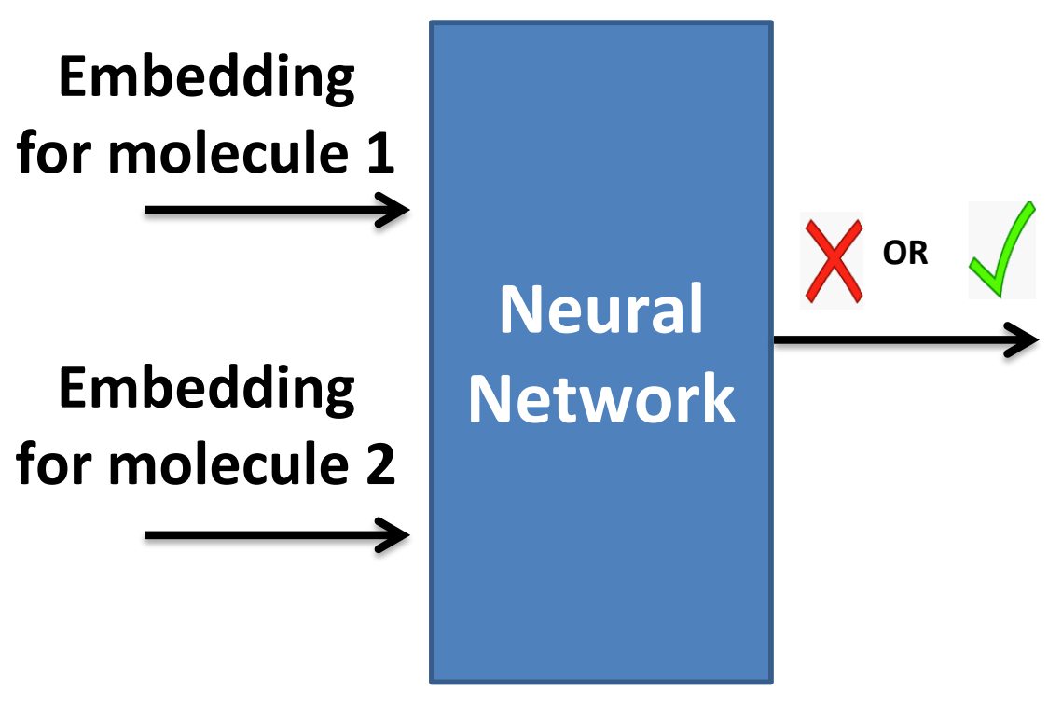
Approach Overview

Graph Embedding



Adapted from [1]

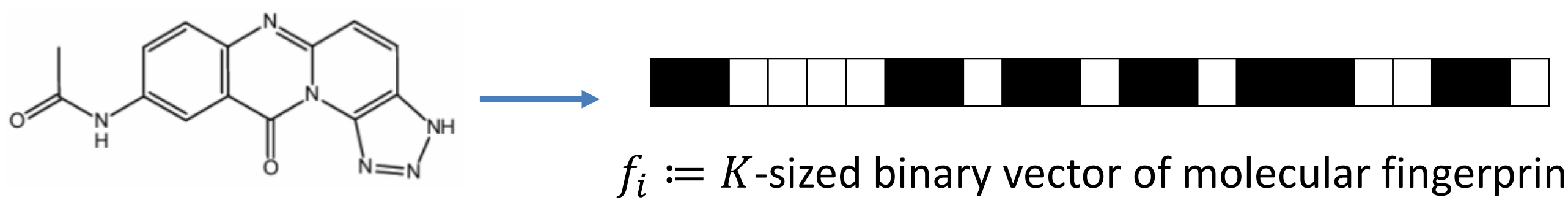
Use learned embeddings to predict likelihood of molecular interactions



Graph Construction

- Use reactions in the KEGG [2] database
  - all reactions are reversible; remove cofactors
- Every molecule is a node
- Each substrate-product pair within a reaction is an undirected edge
- Edge attributes: enzyme commission (EC) number or reaction class (RC)
- Node attributes: fingerprints (MACCS [3] or PubChem [4])

Molecule i

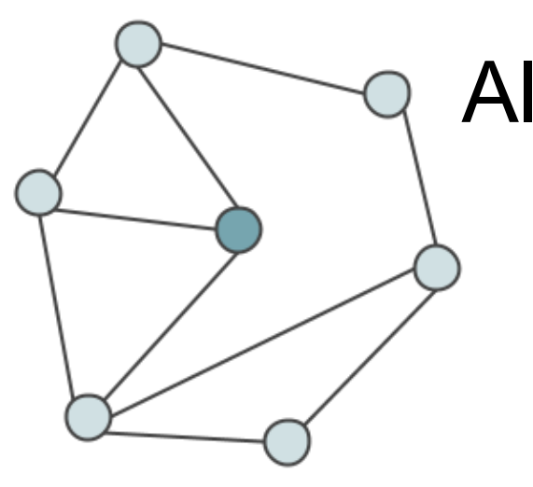


$f_i :=$ K-sized binary vector of molecular fingerprint

Enzymatic Link Prediction (ELP)

1 Embedding Propagation on Graph

We use Embedding Propagation [5], a graph embedding method, to learn embedding vectors of nodes



All embeddings are randomly initialized:

- Connectivity-based node embeddings $\{\mathbf{u}_i\}$,
- Fingerprint embeddings $\{\mathbf{v}_k\}$, one for each fingerprint entry
- Enzyme embeddings $\{\boldsymbol{z}_r\}$, one for each enzyme label

- Fingerprint-based node embeddings $\{\mathbf{u}_i^{fp}\}$ are constructed from fingerprint embeddings

$$\mathbf{u}_i^{fp} = \frac{1}{\sum_{k=1}^{K} f_{ik}} \sum_{k=1}^{K} f_{ik}\mathbf{v}_k$$

k^{th} value of node i's fingerprint

Fingerprint embedding of entry k

- Reconstruct node embedding ($\tilde{\mathbf{u}}_i$) from the embeddings of its neighbors

$$\tilde{\mathbf{u}}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{u}_j + \alpha\, \mathbf{z}_{r(i,j)}$$

Neighbors of node i

Node embedding of node j

Enzyme embedding of the edge (i,j)

- Margin-based ranking loss.
- Aim to maximize the similarity between the *reconstruction of node embedding* $\tilde{\mathbf{u}}_i$ with *node embedding* $\mathbf{u}_i$

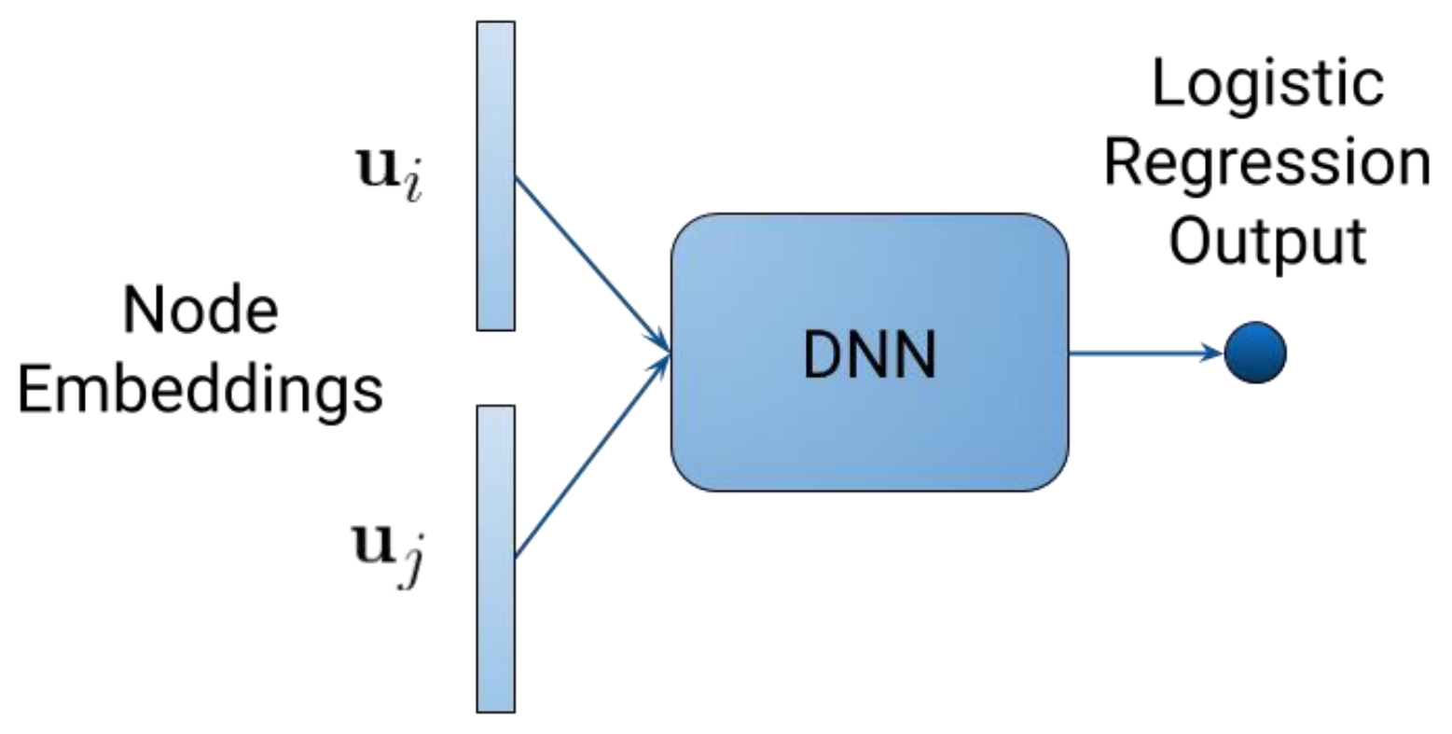
$$\mathcal{L} = \sum_{i \in V} \sum_{j \in V, j \neq i} \max\{\gamma - \tilde{\mathbf{u}}_i^{\top}\mathbf{u}_i + \tilde{\mathbf{u}}_i^{\top}\mathbf{u}_j, 0\}$$

Random node j as the negative example for each node in every iteration

- Concatenate $\mathbf{u}_i$ and $\mathbf{u}_i^{fp}$ to form final node embedding vectors

2 Link Prediction Using Embedding Vectors

Train a logistic regression model using deep neural nets to predict the likelihood of an edge between two nodes



Experiments & Results

Transductive Learning

- Model is trained on all nodes and evaluated for edge recovery on a held out set of test edges.
- Training graph must be connected

Baseline: other embedding methods

Baseline: no connectivity embedding

Method	Connectivity Embedding	Node Attribute	Edge Attribute	AUC Test Ratios 0.1	0.3	0.5
A. Connectivity-based embeddings only						
ELP	Yes	–	–	0.801	**0.789**	0.761
node2vec	Yes	–	–	0.824	0.736	**0.776**
DeepWalk	Yes	–	–	**0.847**	0.763	0.749
B. Connectivity and one additional attribute						
ELP	Yes	MACCS	–	**0.953***	**0.935***	**0.900**
ELP	Yes	PubChem	–	0.891	0.882	0.864
ELP	Yes	–	EC	0.795	0.808	0.810
ELP	Yes	–	RC	0.810	0.798	0.810
C. Connectivity with one node and one edge attribute						
ELP	Yes	MACCS	EC	0.941	**0.933**	**0.922***
ELP	Yes	MACCS	RC	**0.942**	0.929	0.895
ELP	Yes	PubChem	EC	0.892	0.879	0.867
ELP	Yes	PubChem	RC	0.892	0.876	0.859
D. Embedding based on MACCS fingerprints						
ELP	No	MACCS	–	0.931	0.916	0.898
ELP	No	MACCS	EC	**0.940**	**0.925**	**0.913**
ELP	No	MACCS	RC	0.939	0.904	0.896
E. Embeddings based on PubChem fingerprints						
ELP	No	PubChem	–	0.665	**0.709**	0.682
ELP	No	PubChem	EC	**0.745**	0.707	**0.728**
ELP	No	PubChem	RC	0.728	0.706	0.720
F. Jaccard index similarity scoring; no embeddings						
Jaccard	No	MACCS	–	**0.808**	**0.778**	**0.767**
Jaccard	No	PubChem	–	0.542	0.526	0.535

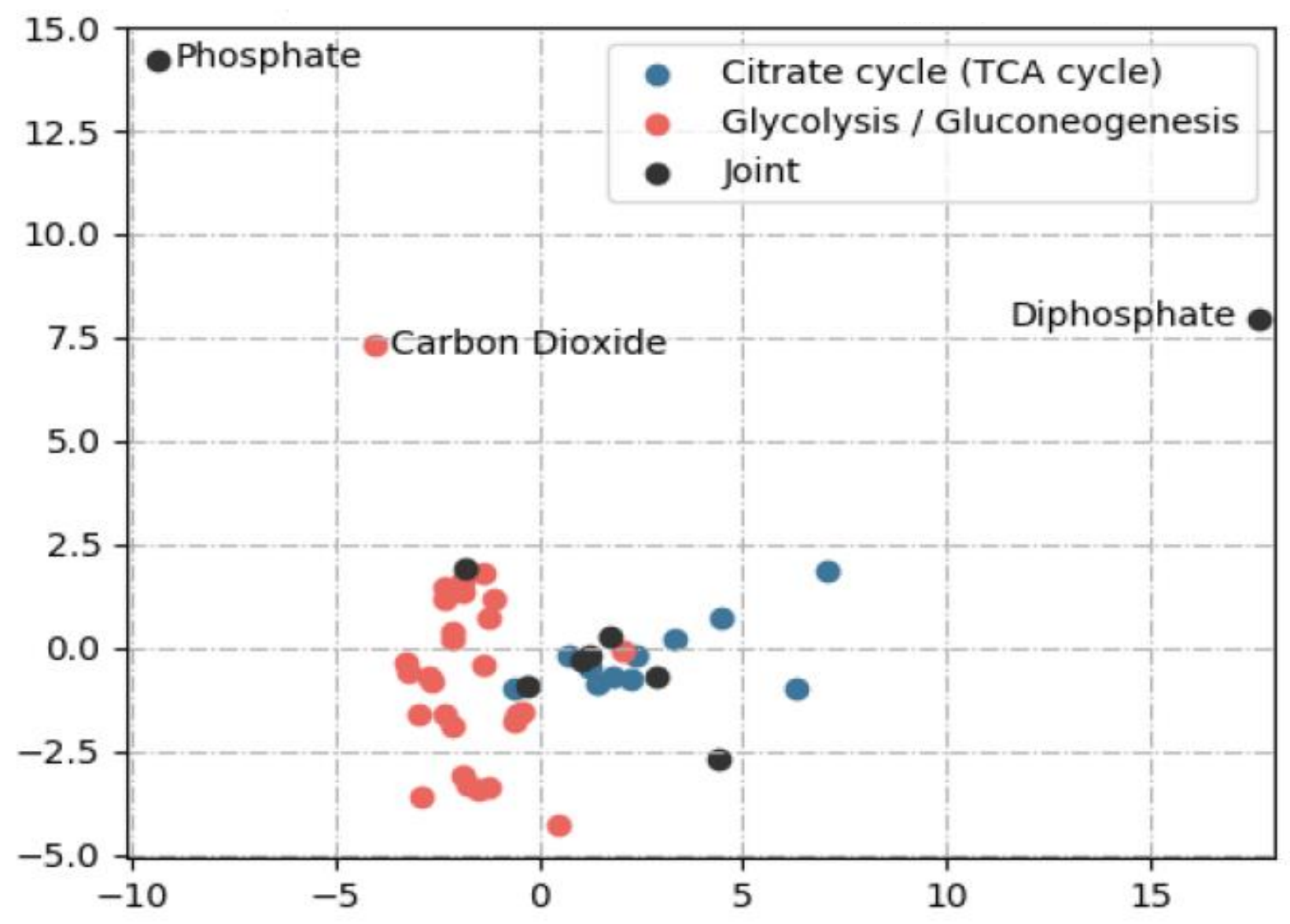
Bold value indicates best result.
* Indicates best overall results

Inductive Learning

- Model is trained to predict possible interactions for *out-of-sample* nodes excluded from training
- This type of prediction is made possible by only using fingerprint-based node embeddings

Method	Connectivity Embedding	Node Attribute	AUC
A. Embeddings based on node attributes			
ELP	Yes	MACCS	**0.921**
ELP	Yes	PubChem	0.605
B. Jaccard index similarity scoring			
Jaccard	No	MACCS	0.744
Jaccard	No	PubChem	0.553

Other applications of embeddings: Visualization of Metabolites within Pathways using t-SNE



Conclusion

ELP learns molecular representations that capture graph connectivity, enzymatic properties, and structural molecular properties

- ELP shows high accuracy in link prediction when using both graph connectivity and molecular attributes
- ELP can be used as a guide to identifying catalyzing enzymes when constructing novel synthesis pathways or predicting interaction between microbes and human hosts
- ELP can enhance link prediction in chemical networks, where previously rule-based and path-based link prediction respectively yielded 52.7% and 67.5% prediction accuracy [6]

References

1. Cai, H., Zheng, V. W., & Chang, K. C. C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 30(9), 1616-1637
2. Kanehisa, M., Sato,Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, **44**(D1), D457–D462.
3. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, **42**(6), 1273–1280.
4. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. (2015). Pubchem substance and compounddatabases. *Nucleic acids research*, **44**(D1), D1202–D1213.
5. Duran, A. G. and Niepert, M. (2017). Learning graph representations with embedding propagation. In *Advances in neural information processing systems*, pages 5119–5130.
6. Segler, M. H., & Waller, M. P. (2017). Modelling chemical reasoning to predict and invent reactions. Chemistry–A European Journal, 23(25), 6118-6128.

Acknowledgements

This research is supported by NSF under Award Number CCF-1909536 and also by NIGMS of the National Institutes of Health under Award Number R01GM132391. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.