Organizing Chemical Reactions with Network Representation Learning

Blake B. Gaines[1], Minghu Song[2], Jinbo Bi[1]
[1]UConn Computer Science and Engineering, [2]UConn Biomedical Engineering

Abstract

For both chemists and computers, datasets of chemical reactions are hard to use. One approach is to represent reactions as points in space, where the distance between the points corresponds to the similarity between reactions. We do this by first converting reactions into nodes in a network of chemical data, and then test two existing techniques for converting network nodes into points.

Creating a Network of Reactions

Figure 1: An example reaction (top) and it's template (bottom)

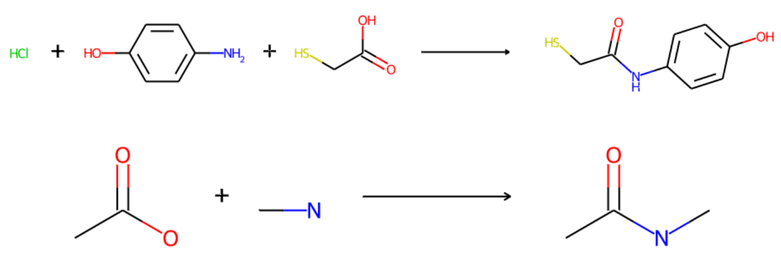


Figure 2: The overall structure

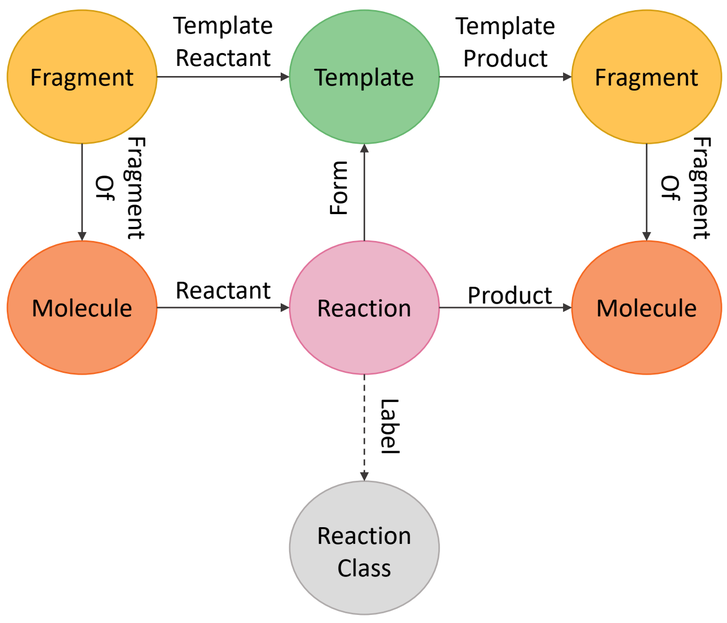
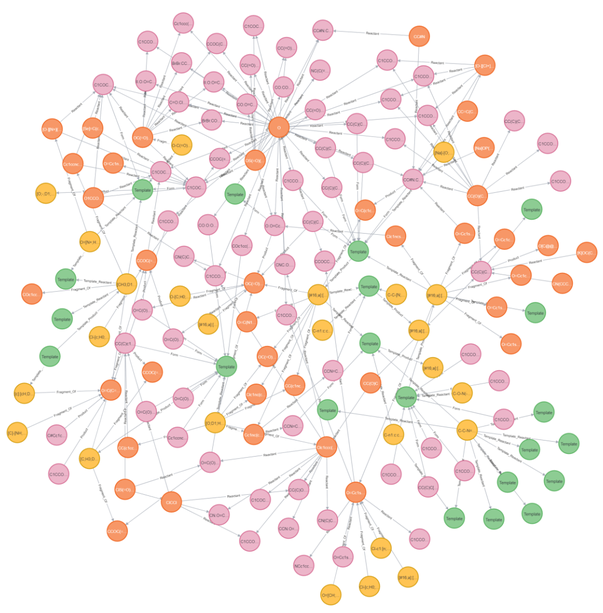


Figure 3: Part of the actual network



Turning Reaction Nodes to Points in Space

General Procedure

- Randomly Place Points
- Repeat:
  - Pick a small group of nodes
  - Check how they are related in the network
  - Check how their points are arranged
  - Move the points slightly to better fit the relationships

Figure 5: Geometric relationships enforced by RotatE for relationships of the form (h) -r-> (t).

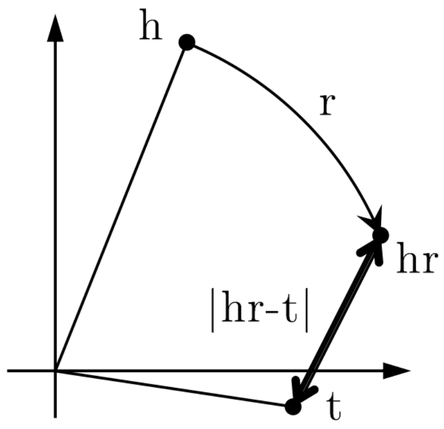


Figure taken from the original RotatE Paper[4]

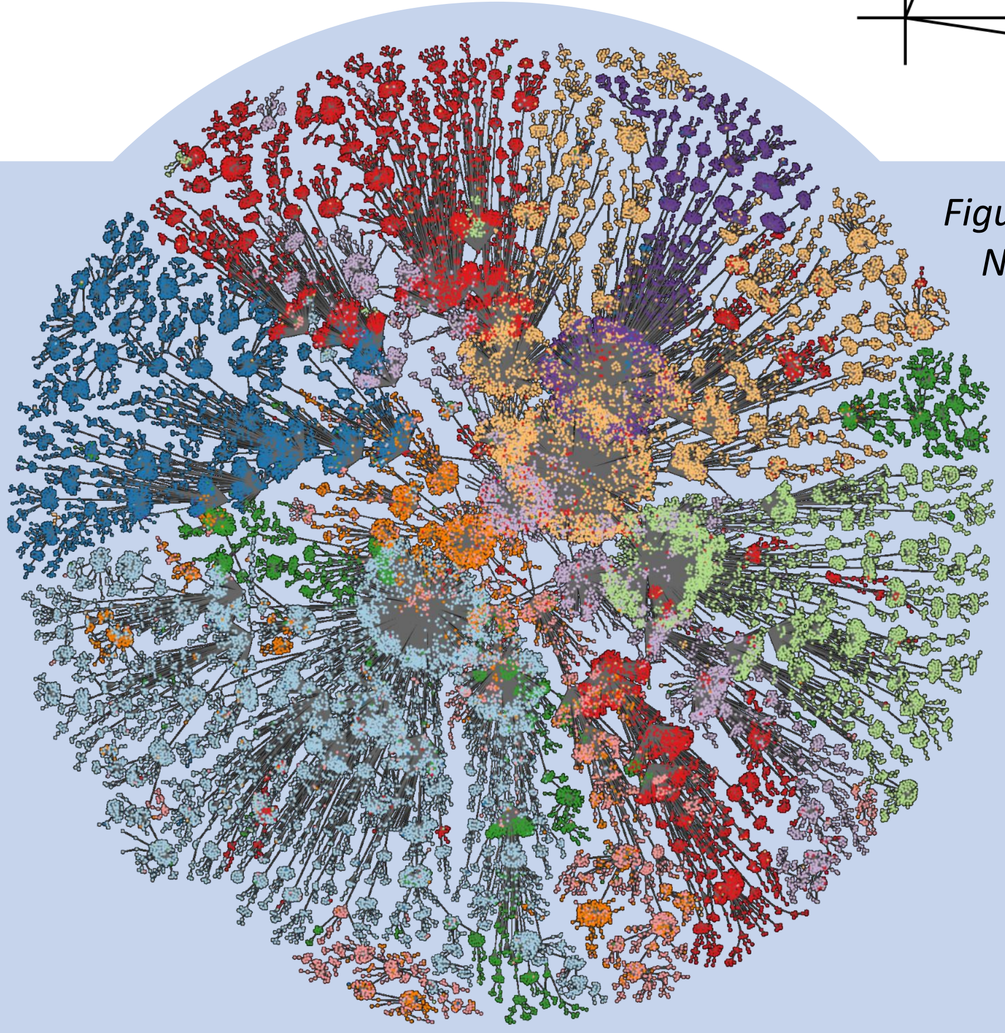


Figure 7: TMAP plot of Node2Vec-based vectors for "Heteroatom Alkylation and Arylation" reactions

Legend of Reaction Subclasses

- Chloro- N-arylation
- SNAr ether synthesis
- Aldehyde reductive amination
- Fluoro N-arylation
- Bromo N-arylation
- Bromo N-alkylation
- Williamson ether synthesis
- Bromo Buchwald-Hartwig amination
- Chloro N-alkylation
- Mitsunobu Aryl ether synthesis

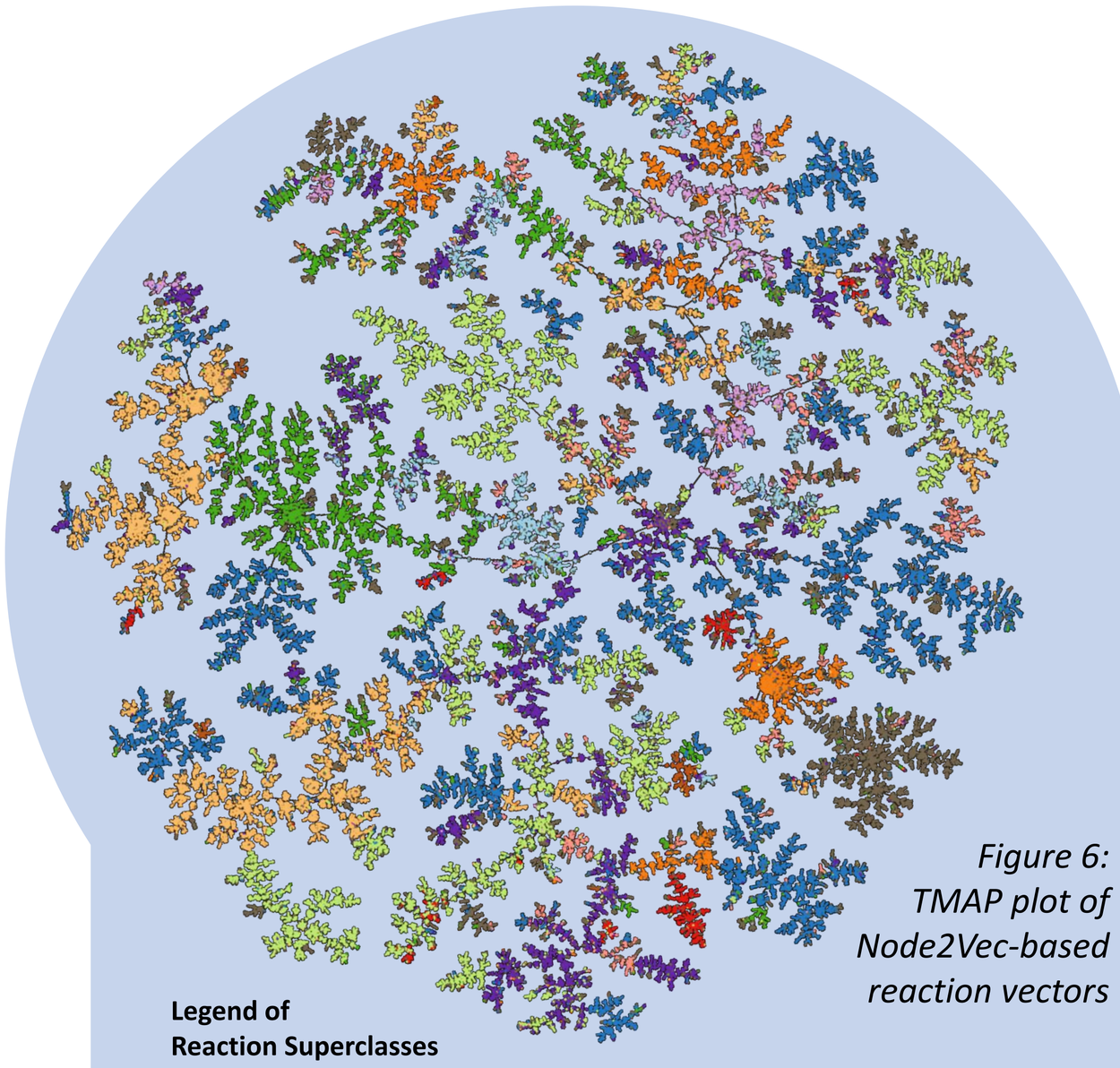


Figure 6: TMAP plot of Node2Vec-based reaction vectors

Legend of Reaction Superclasses

- Unrecognized
- Heteroatom Alkylation and Arylation
- Acylation and Related Processes
- C-C Bond Formation
- Heterocycle Formation
- Protections
- Deprotections
- Reductions
- Oxidations
- Functional Group Interconversions
- Functional Group Additions
- Resolutions

Predicting Reaction Labels

Table 1: Classification Performance for Unsupervised Representations

Fingerprint	Classifier	Accuracy	Overall MCC	Overall CEN
Traditional	5-NN	0.688	0.675	0.218
	LR	0.671	0.651	0.199
Machine Learning	5-NN	0.811	0.802	0.126
	LR	0.781	0.768	0.141
Ours (Node2Vec)	5-NN	0.815	0.809	0.114
	LR	**0.946**	**0.944**	**0.039**

Table 2: Classification Performance for Supervised Representations

Fingerprint	Classifier	Accuracy	Overall MCC	Overall CEN
Machine Learning	**5-NN**	**0.986**	**0.986**	**0.012**
	LR	0.986	0.985	0.012
Ours (RotatE)	5-NN	0.843	0.841	0.102
	LR	0.754	0.755	0.158
	geometric	0.696	0.701	0.173
Ours (Node2Vec)	5-NN	0.851	0.850	0.088
	LR	0.962	0.960	0.031

Traditional: Morgan Fingerprints
Machine Learning: RXNFP[3] (SMILES-Based Transformer)

Summary

- Networks of chemical reactions naturally encode patterns in their structure
- Existing methods can effectively encode this structure in vector representations
- Without supervision, this method outperforms both traditional and ML-based embeddings
- With supervision, this method achieves competitive performance with a fraction of the cost

References

1. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., & Jensen, K. F. (2017). Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science*, *3*(5), 434–443.
2. Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *ArXiv:1607.00653 [Cs, Stat]*.
3. Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., & Reymond, J.-L. (2021). Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, *3*(2), 144–152.
4. Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *ArXiv:1902.10197 [Cs, Stat]*.