Multimodal Molecule Reaction Mining from Knowledge Graphs

Carl Edwards
cne2@illinois.edu
University of Illinois Urbana-Champaign
Urbana-Champaign, Illinois, USA

ABSTRACT

Many new and exciting chemicals have been discovered. However, the reactions to produce them often produce low reaction yields or require expensive intermediate reactants. The chemicals cannot be used because they are too difficult to create at the large scales required for industrial use. In order to remedy this problem, finding new reactions with improved yields or scalability is very desirable. I plan to use a knowledge graph of chemical reactions and molecule information to perform knowledge graph completion to find new reactions. In particular, I plan to approach this problem using joint embedding techniques.

KEYWORDS

molecules, reactions, knowledge graphs, embeddings

ACM Reference Format:
Carl Edwards. 2021. Multimodal Molecule Reaction Mining from Knowledge Graphs. In *Proceedings of CS 512 (Data Mining Principles).* , 2 pages.

1 INTRODUCTION

Chemicals are ubiquitous in everyday life. Whether enabling the latest breakthroughs and products in technology, medicine, or materials, they are an essential component of modern day life. In order to create the chemicals, reactions consisting of other chemicals are required. However, finding reaction routes which are scalable and efficient can be very challenging. For example, a critical drug might be found for some disease, but there is no way to produce it at a large enough scale to be made commercially. To address this issue, it is important to find new reactions to create existing compounds. To do this, I propose using a joint-embedding based method between molecules and reactions framed as a knowledge graph.

2 RELATED WORKS

In the past, chemists would consider hundreds of possible reactants and conduct many experiments to find new reactions. However, the recent success of data mining and machine learning techniques opens a new avenue to find these reactions. Work such as [7, 12] focus on using computers to plan reactions. [6] focuses on automating reaction network construction from reaction databases. Additionally, recent approaches such as [1, 8, 11] apply deep learning to plan retrosynthetic reactions or drug-drug interactions. In particular, [10] creates a knowledge graph from Reaxys [1] data. The data consists of 14.4 million molecules and 8.2 million binary reactions between them, which consists of the majority of chemical reactions in the literature up to 2013. They use reactions from 2014 and 2015 as a validation dataset and find that they can independently re-discover 35% of challenging novel reactions discoverd in those years (reactions without 'template' reactions from prior work). This work uses a breadth-first search based model which prunes possible reaction paths based on Tanimoto similarity of reaction fingerprints.

Additionally, machine learning has become more popular in cheminformatics. Representations for molecules such as mol2vec [3] and MolBert [2] create embedding representations by leveraging the recent success from natural language processing algorithms. These are applicable to various downstream tasks, such as molecule property prediction.

3 DATASETS

Unfortunately, the chemistry community is much more closed and proprietary than data mining. There are not standard datasets and many of the best databases are proprietary. For example, I would like to use the knowledge graph created in [10], but it was created using Reaxys, a commercial database. However, UIUC appears to have access to Reaxys [2]. I plan to follow the methodology from [10] to create a similar reaction database, although I might restrict its size.

If using reaxys does not work, several other potential options are listed at http://organicworldwide.sitehosting.be/content/reaction-databases. Additionally, NIST [5] has a kinetics database of reactions which could prove useful. Finally, the PubChem [4] dataset contains information on millions of chemicals and a RDF with links between some of them. Although these links are not necessarily reactions, predicting relations in this case would also be useful. This graph, however, may prove to be too sparse. Further, PubChem has about 16,000 links to molecules in WikiData. This could also potentially serve as a useful dataset for infusing knowledge graph embeddings with chemical information.

4 PLAN AND TIMELINE

For my project, I propose using molecule representations to help 'steer' the embedding of a reaction knowledge graph to improve performance on relation prediction. Rather than use a domain-engineered approach, I hope to use and end-to-end model for link (reaction) prediction. I plan to incorporate the molecule embeddings in two ways. First, I plan to use a multi-layer perceptron to incorporate the information after the node embedding. Second, I plan to create a separate, aligned embedding space. I have been working on embedding text in this manner, and I believe replacing it with knowledge graphs would be a promising alternative which

[1] https://www.reaxys.com/

[2] https://www.library.illinois.edu/chx/reaxys/

Data Mining Principles, Spring 2021, Illinois, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

is well-suited to this class due to its focus on graph and network mining this semester.

- March 25th: Create dataset
- April 7th: Create baseline using a KG embedding model such as relational GCN [9]
- May 1st: Incorporate molecule representation into KG embeddings.

REFERENCES

[1] Javier L Baylon, Nicholas A Cilfone, Jeffrey R Gulcher, and Thomas W Chittenden. 2019. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *Journal of chemical information and modeling* 59, 2 (2019), 673–688.

[2] Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230* (2020).

[3] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58, 1 (2018), 27–35.

[4] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* 47, D1 (2019), D1102–D1109.

[5] JA Manion. 2008. NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.4. 3, Data version 2008.12. *http://kinetics. nist. gov/* (2008).

[6] Pieter P Plehiers, Guy B Marin, Christian V Stevens, and Kevin M Van Geem. 2018. Automated reaction database and reaction network analysis: extraction of reaction templates using cheminformatics. *Journal of cheminformatics* 10, 1 (2018), 1–18.

[7] Yannick Pouliot, Annie P Chiang, and Atul J Butte. 2011. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clinical Pharmacology & Therapeutics* 90, 1 (2011), 90–99.

[8] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences* 115, 18 (2018), E4304–E4311.

[9] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[10] Marwin HS Segler and Mark P Waller. 2017. Modelling chemical reasoning to predict and invent reactions. *Chemistry–A European Journal* 23, 25 (2017), 6118–6128.

[11] Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. 2020. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning*. PMLR, 8818–8827.

[12] Ivar Ugi, Johannes Bauer, Klemens Bley, Alf Dengler, Andreas Dietz, Eric Fontain, Bernhard Gruber, Rainer Herges, Michael Knauer, Klaus Reitsam, et al. 1993. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angewandte Chemie International Edition in English* 32, 2 (1993), 201–227.