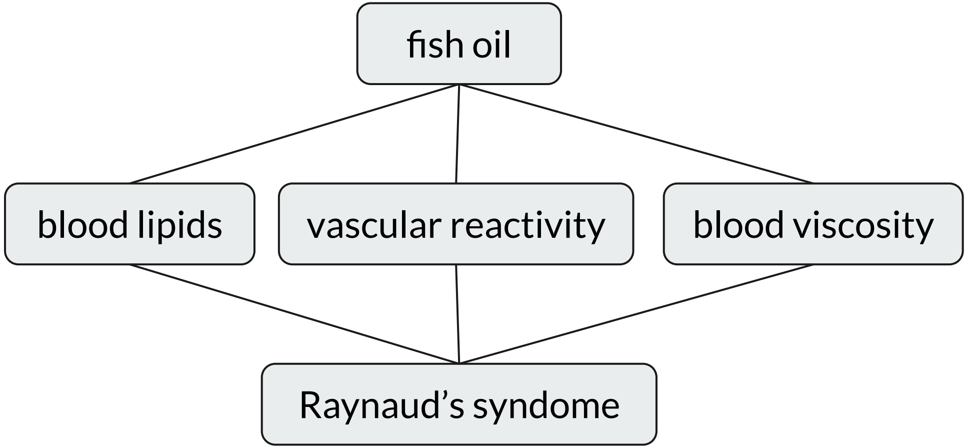
Integrating publication signals for literature-based knowledge discovery

Undiscovered public knowledge

Paper from 1986 suggested that the knowledge of a treatment for a disease may already be in the literature

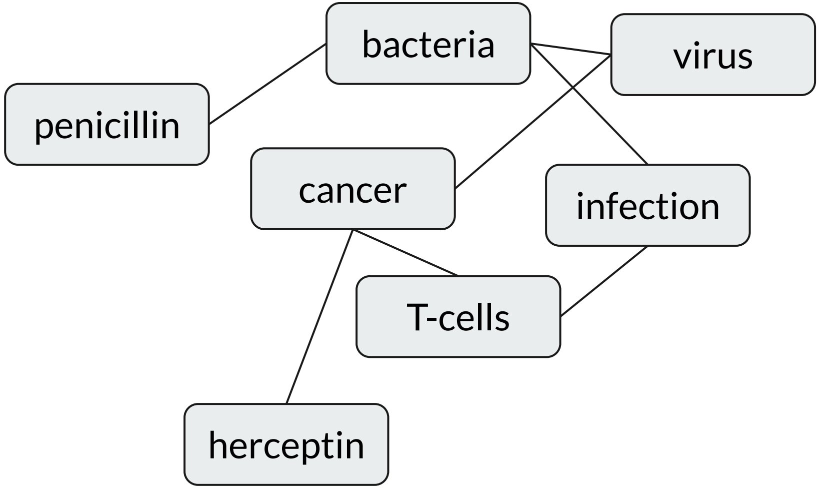
BUT: The right links may not have been made

Literature-based knowledge discovery (LBD) is the challenge to find those links!



Swanson, Don R. "Fish oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in biology and medicine* 30.1 (1986): 7-18.

A co-occurrence knowledge graph

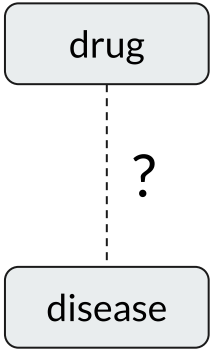


LBD often focuses on co-occurrences:

- Two biomedical entities (e.g. drug, disease, etc) appearing together
- Could be within a sentence, paper title, abstract or whole paper

Link predictions on a knowledge graph

- Literature based knowledge discovery involves predicting new links in this graph
- Link prediction is a frequent problem with applications like predicting friends on a social network

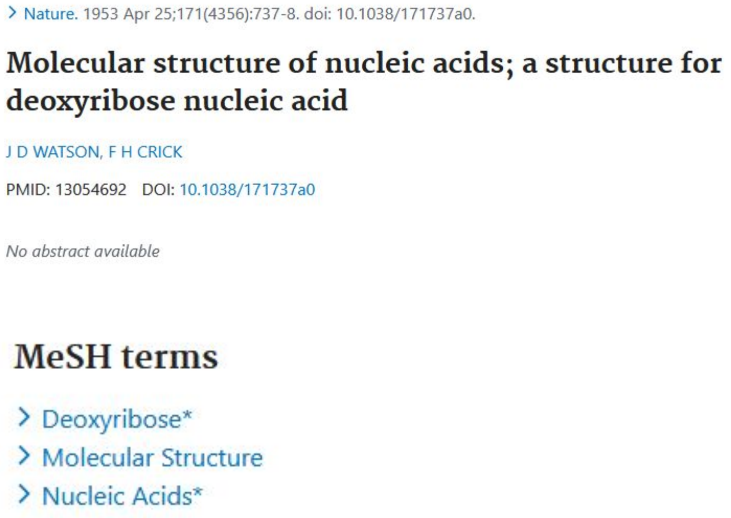


Medical subject headings (MeSH)

Where can we get a co-occurrence network? (with extra metadata that we might want about the source of the knowledge)

PubMed is “the” repository of abstracts of biomedical research publications

Almost all articles are tagged with the topics that they discuss. These are MeSH terms



Nature. 1953 Apr 25;171(4356):737-8. doi: 10.1038/171737a0.

Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

J D WATSON, F H CRICK

PMID: 13054692 DOI: 10.1038/171737a0

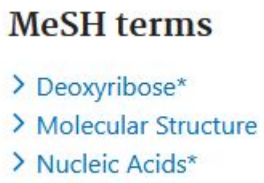
No abstract available

MeSH terms

- Deoxyribose*
- Molecular Structure
- Nucleic Acids*

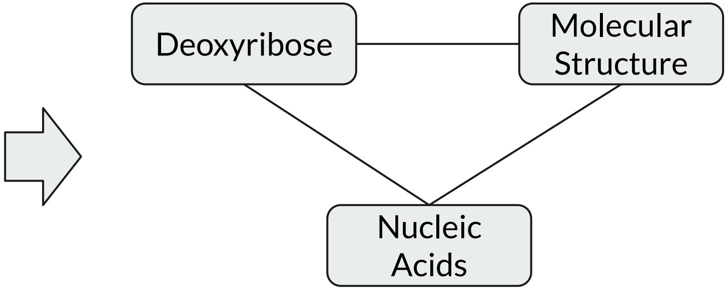
https://pubmed.ncbi.nlm.nih.gov/13054692/

Creating a knowledge graph from MeSH



MeSH terms

- Deoxyribose*
- Molecular Structure
- Nucleic Acids*

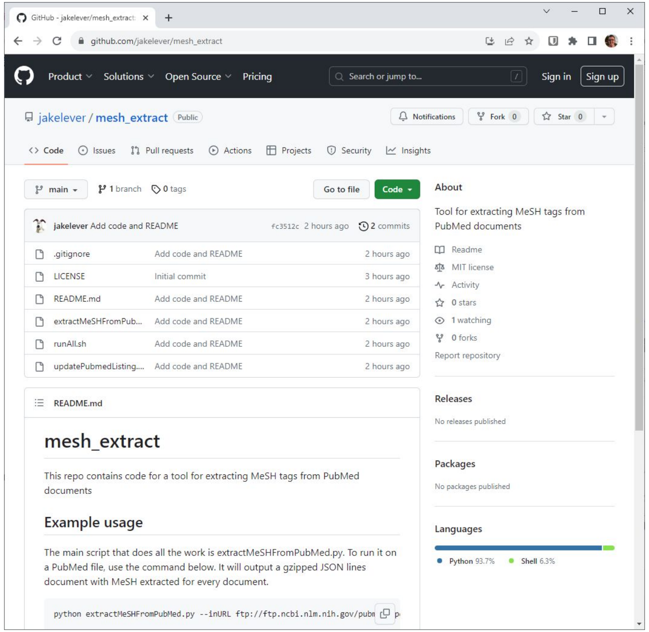


- Create all pairs of MeSH terms that appear in the same document
- Do it for a big set of publications
- Keep track of the date a co-occurrence first appears
  - And any other useful data

Getting MeSH for all of PubMed

I can provide a MeSH dataset generated from the bulk downloaded PubMed

https://github.com/jakelever/mesh_extract



jakelever Add code and README fc3512c 2 hours ago 2 commits

.gitignore	Add code and README	2 hours ago
LICENSE	Initial commit	3 hours ago
README.md	Add code and README	2 hours ago
extractMeSHFromPub...	Add code and README	2 hours ago
runAll.sh	Add code and README	2 hours ago
updatePubmedListing....	Add code and README	2 hours ago

Tool for extracting MeSH tags from PubMed documents

mesh_extract

This repo contains code for a tool for extracting MeSH tags from PubMed documents

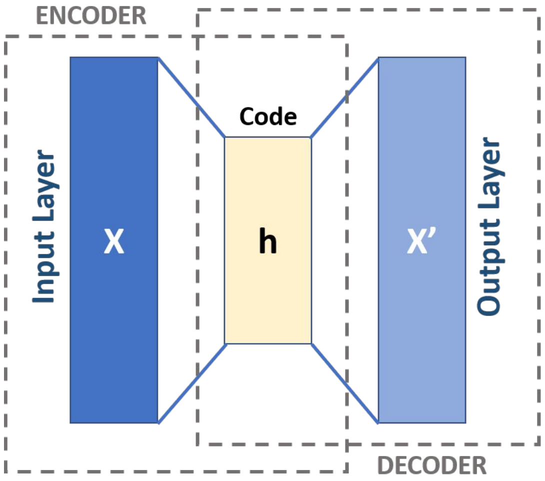
Example usage

The main script that does all the work is extractMeSHFromPubMed.py. To run it on a PubMed file, use the command below. It will output a gzipped JSON lines document with MeSH extracted for every document.

```
python extractMeSHFromPubMed.py --inURL ftp://ftp.ncbi.nlm.nih.gov/pubm
```

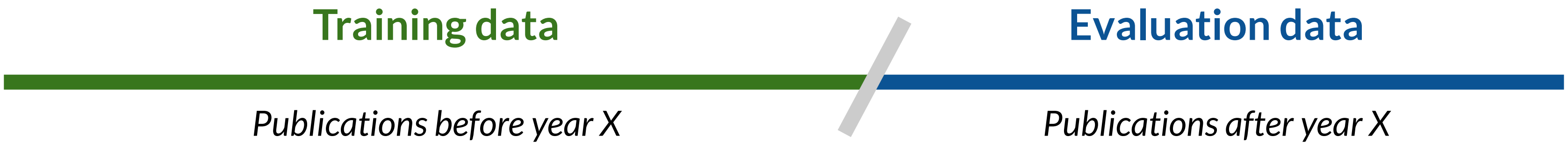
Existing methods for predicting links

- Methods from recommendation systems
  - Matrix decomposition
- Deep learning methods
  - Auto-encoder
  - Graph neural networks
  - Transformer-based methods
- Graph traversal methods
  - Simpler A-B-C methods to find shared links
- The choice is yours



https://en.wikipedia.org/wiki/Autoencoder#/media/File:Autoencoder_schema.png

Evaluating by splitting in time



Training data

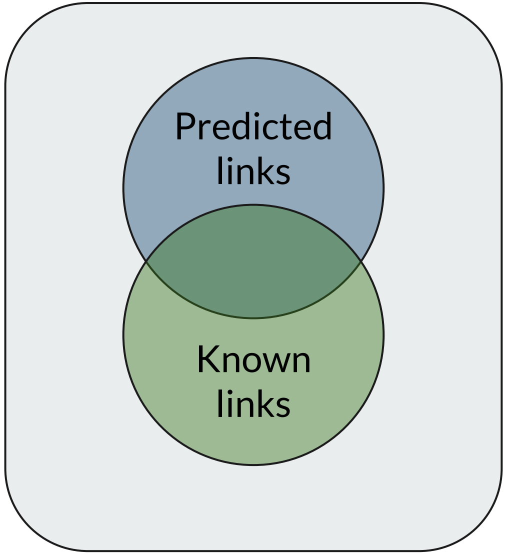
Evaluation data

Publications before year X

Publications after year X

- Build a co-occurrence graph up to a certain year (e.g. 2015)
- Make predictions on the co-occurrence graph
- Track all new co-occurrences in publications after that year and compare with the predictions

Evaluation metrics



Treating as classification

Is this classification or ranking?

- Classification methods:
  - Precision, recall, F1, etc
- Ranking
  - Hits @ 10, Mean Reciprocal Rank, etc

LBD is typically treated as a classification task

Note: This problem has an odd property:

- A link not existing between A and B may be missing or actually negative

Top predicted links connected to "Alzheimer's"

1. ABCB1 gene
2. cancer
3. amyloid [correct]
4. aspirin
5. horseradish

Treating as ranking

Possible research ideas

- Does the popularity of a paper (and associated co-occurrences) provide a useful signal for predictions?
    - Source data: https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collection_/4586573/44
- Does the date of publications of co-occurrences provide a useful signal?
    - Perhaps older information is more or less useful?
    - Perhaps information that was only talked about in the 1990s is not helpful
- Your own ideas?

Some papers to get you started

A collaborative filtering-based approach to biomedical knowledge discovery

Graph embedding-based link prediction for literature-based discovery in Alzheimer’s Disease

Discovering and visualizing indirect associations between biomedical concepts