# Presentation about the Research Paper: Constructing knowledge graphs and their biomedical applications

*by* **Lennard Feuerbach**

*on the* **19th of November, 2024**

# Table of Contents

# 1. Overview of Paper Metadata

- By David N. Nicholson and Casey S. Greene from UPenn in the U.S.
- Department of Systems Pharmacology and Translational Therapeutics
- Released in Computational And Structural Biotechnology Journal
- Article uploaded on 12 March 2020 and accepted on 23 May 2020
- Licensed under the Creative Commons BY 4.0 (with attribution to author)
- Funding by Gordon and Betty Moore Foundation and National Institutes of Health

# 2. Key Points from the Abstract

- Knowledge graphs can represent biomedical concepts and relationships
- Traditional construction of biomedical graphs relied on expert-curated databases
- The creation is increasingly adopted and done by automated systems
- This paper focuses on knowledge graph construction, techniques and applications
- Advances in machine learning for biomedicine are opening new opportunities

# 3. Introduction and Background

- In biomedicine, graphs been used in e.g. gene prioritization or drug repurposing
- Defining a knowledge graph is challenging due to conflicting definitions
- Biomedical knowledge graphs are integrating expert-derived information into graphs
- Relationships in these graphs are typically unidirectional, also can be bidirectional

# 4. Building Biomedical Knowledge Graphs

- Knowledge graphs can be built from existing databases or text resources
- Databases usually created by domain experts with manual or automated methods
- Manual curation requires reading and annotating papers to find relationships
- Automated approaches use *ML* or *NLP* to quickly identify relevant sentences
- In the following approaches are discussed, with their strengths and weaknesses

## 4.1 Constructing Databases and Manual Curation

- Database construction began 1956 with protein sequence database for insulin
- Constructing databases involves curators extracting relationships from texts

| Database | Number of Entries | [...] | Method of Population |
|---|---|---|---|
| BioGrid | 572,084 | | Manual and Automated Curation |
| COSMIC | 35,946,704 | | Manual Curation |
| UniProt | 560,823 | | Manual and Automated Curation |

**Figure 1:** Databases constructed after the principles (according to [1, p. 1416])

# 4. Building Biomedical Knowledge Graphs

## 4.1 Constructing Databases and Manual Curation

- Manually curated databases are precise but suffer from rapid publication rates
- Semi-automatic methods accelerate curation by pre-filtering irrelevant sentences
- Automated systems excel at finding common relationships, struggle with rare ones
- Manual curation remains essential for producing gold standard datasets
- Future databases should use automated methods for extraction and manual curation

## 4.2 Text Mining for Relationship Extraction

### 4.2.1 Rule-based Relationship Extraction

- Uses keywords and grammatical patterns to identify relationships in text
- Keywords are derived from expert knowledge or pre-existing ontologies
- Grammatical patterns are constructed via experts curating parse trees
- Rule-based methods require manual effort and expert knowledge

## 4.2 Text Mining for Relationship Extraction

### 4.2.1 Rule-based Relationship Extraction

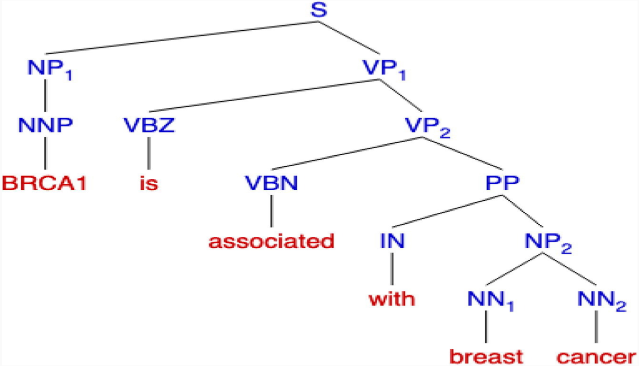


**Figure 2:** Constituency parse tree [1, p. 1417]

## 4.2 Text Mining for Relationship Extraction

### 4.2.2 Extracting Relationships without Labels

- Unsupervised extraction methods infer associations from text without labels
- These methods often involve clustering or statistical calculations, like co-occurrence
- Databases like *DISEASES* and *STRING* used co-occurrence on PubMed abstracts
- Unsupervised methods enable rapid relationship extraction without annotations
- Full-text mining and sentence simplification hold potential to further improvement

## 4.2 Text Mining for Relationship Extraction

### 4.2.3 Supervised Relationship Extraction

- Supervised extraction uses labels to distinguish positive from negative examples
- Pre-labeled datasets constructed as gold standards to support these methods
- Approaches that use these datasets include linear and non-linear classifiers
- Linear classifiers include *SVMs*, while non-linear include Deep Learning techniques
- Semi-supervised and weak supervision enhance the extraction for *ML* classifiers

# 5. Applying Graphs to Biomedical Challenges

## 5.1 Unifying Representational Learning Techniques

- Mapping high into a low dimensional space improves modeling performance
- Knowledge graphs are represented as dense vectors in low-dimensional space
- Once this space has been constructed, *ML* techniques can work with them
- We group techniques that construct this space into the following three categories:
  - Matrix factorization, various techniques that use linear algebra
  - Translational distance models, treat edges in graphs as linear transformations
  - Neural network models, *ML* models for non-linear transformations

# 5. Applying Graphs to Biomedical Challenges

## 5.2 Unifying Applications

- Knowledge graphs are applied to different biomedical challenges
- **Multi-omic applications**, use graphs to understand biological systems
- **Pharmaceutical applications**, use graph to identify new properties of drugs
- **Clinical applications**, use analyses of these graphs to aid patient care

# 5. Applying Graphs to Biomedical Challenges

## 5.2 Unifying Applications

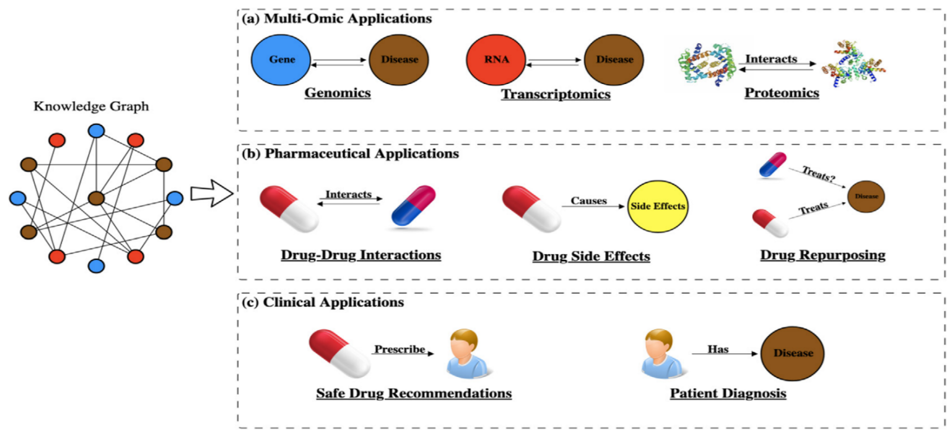


**Figure 3:** Biomedical applications of knowledge graphs [1, p. 1420]

# 6. Conclusion and Final Takeaways

- Knowledge graphs are growing in use in biomedicine and will continue to expand
- Currently, most graphs are built from databases that are manually curated
- There are emerging automated approaches that can support manual curation
- Representing graphs in low-dimensional space can support various *ML* analyses
- *ML* is expected to help uncover new insights from these knowledge graphs

# References


- [1] D. N. Nicholson and C. S. Greene, “Constructing knowledge graphs and their biomedical applications,” Computational and Structural Biotechnology Journal, vol. 18, Elsevier, 2020, pp. 1414-1428, doi: 10.1016/j.csbj.2020.05.017.
- [2] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, . . . P. J. Campbell, “COSMIC: somatic cancer genetics at high-resolution,” Nucleic Acids Research, 2016, doi: 10.1093/nar/gkw1121. PMID: 27899578, PMCID: PMC5210583.
- [3] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, “A survey on knowledge graph embedding: Approaches, applications and benchmarks,” Electronics, vol. 9, no. 5, 2020, p. 750, doi: 10.3390/electronics9050750.