# Extracting Data to Build and Analyze Networks: Integrating Biomedical Entities for Multi-Layer Analysis

Daisuke Terauchi, Ibuki Yasuda, Jeremy Duncan, Muhammad Omar Salah Ud Din
Information Systems Science and Engineering
Intelligent Computer Entertainment Lab
Ritsumeikan University, Japan

***Abstract*—This study presents the design and analysis of a comprehensive knowledge graph (KG) integrating information regarding cancer diseases, symptoms, treatments, and genes. Data were collected automatically from Wikipedia and MedlinePlus via API-based crawling and processed through a structured LLM-driven entity-relationship extraction pipeline. The resulting graph contains over 1,300 nodes and 1,600 relationships, representing a unified view of oncological knowledge. Analyses including connectivity, community detection, centrality, link prediction, and graph traversal reveal highly connected biological patterns: *TP53*, *BRCA1/2*, and *chemotherapy* emerge as central hubs. The graph's modular structure mirrors established cancer taxonomies, confirming the viability of automated biomedical KGs for exploratory research and hypothesis generation.**

***Index Terms*—Knowledge Graph, Cancer, Graph Analysis, Community Detection, Biomedical Informatics, LLM**

## I. Introduction

### A. Problem Description and Motivation

Biomedical data are fragmented across thousands of web resources, each describing diseases, genes, and therapies in diverse textual formats. Researchers and clinicians face challenges integrating this information into unified computational models for inference, prediction, and discovery. A **knowledge graph (KG)** can bridge this gap by representing heterogeneous biomedical concepts as nodes and their relationships as edges, enabling multi-layer network analyses that reveal hidden connections among diseases, molecular mechanisms, and clinical outcomes.

### B. Objective and Scope

The objective of this project is to build a multi-source knowledge graph of cancer-related diseases and conduct a suite of graph-theoretic analyses to uncover structural and biological insights. The scope includes:

- Focus on distinct **cancer types** (approximately 150 diseases).
- Entities include **diseases, genes, symptoms, treatments, diagnoses, and subtypes**.
- Analyses emphasize **graph topology, centrality, and predictive link discovery**.

### C. Data Sources

Two authoritative biomedical repositories were used:

1) **Wikipedia API**: structured, human-readable summaries of diseases.
2) **MedlinePlus Web Service**: curated medical data provided by the U.S. National Library of Medicine.

## II. Methodology

### A. Data Source Description

The KG pipeline ingests semi-structured textual data; Wikipedia entries provide user-moderated summaries, causes, and subtypes, while MedlinePlus provides clinically verified symptoms, diagnoses, and treatments.

### B. Data Acquisition

A custom Python crawler automates acquisition of raw data. Each record is saved with provenance metadata. Wikipedia is accessed via its REST API for plaintext extracts, while MedlinePlus is accessed through its XML search API. MedlinePlus search results are ranked using heuristics to weight title, `altTitle`, and `FullSummary` similarity to the disease name.

### C. Cleaning and Preprocessing

HTML tags and boilerplate sections (e.g., "Start Here") are removed, and text is normalized for large-language-model (LLM) processing. This ensures compatibility with token constraints while maintaining biomedical accuracy.

### D. Entity and Relationship Extraction

Using Google AI Studio API (gemini-2.5-flash-lite), structured entities and relations are extracted following this schema:

```
{
 "disease_name": "",
 "synonyms": [],
 "summary": "",
 "causes": [],
 "risk_factors": [],
 "symptoms": [],
 "diagnosis": [],
 "treatments": [],
```

```
 "related_genes": [],
 "subtypes": []
}
```

Validated JSON outputs are merged into a unified dataset.

### E. Graph Construction

The graph is built using **NetworkX**, where each entity type corresponds to a node label and each relation forms a typed edge (e.g., `has_symptom`, `associated_gene`, `treated_by`). Interactive visualization is provided using **PyVis**.

### F. Tools and Frameworks

Python 3.12, NetworkX, PyVis, Pandas, and Leiden/Louvain community-detection algorithms were employed. Adamic–Adar, Jaccard, and Preferential Attachment were used for link prediction, while ChatGPT-5 supported interpretive analysis.

### G. Data Definition

The cancer knowledge graph represents entities across multiple biomedical layers, including *diseases, genes, treatments, symptoms, risk factors, biomarkers, and biological pathways*. Each entity type corresponds to a node label, while their interactions (e.g., *associated_gene*, *treated_by*, *has_symptom*) form typed edges. This schema allows both molecular and clinical aspects of oncological data to coexist within a unified graph, enabling cross-domain reasoning and multi-scale analysis.

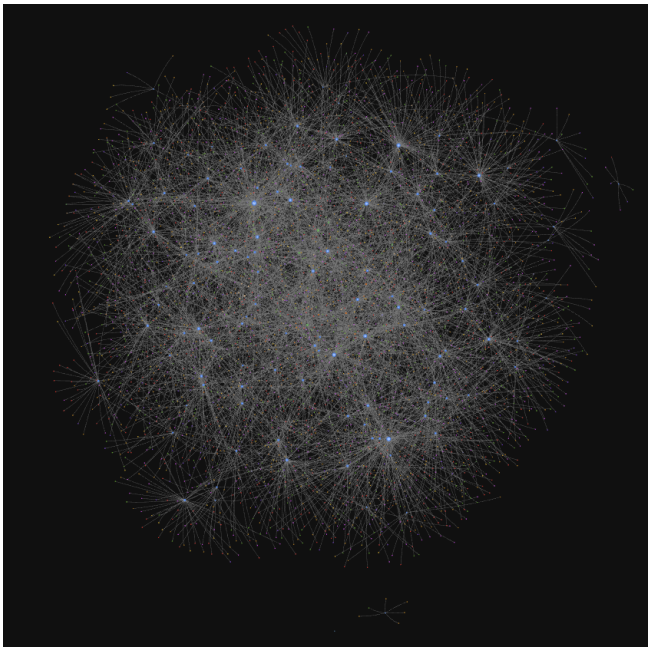

Fig. 1: Full cancer knowledge graph visualization.

## III. Graph Analysis and Experiments

### A. Data Justification

The chosen entity and relationship schema follows design principles from established biomedical knowledge graphs such as *Hetionet* [**?**] and *DisGeNET* [**?**], where disease–gene–drug triplets form the foundation for mechanistic and therapeutic inference. Integrating additional layers—such as *symptoms, biomarkers,* and *risk factors*—follows recommendations from Kaur *et al.* [**?**], emphasizing multi-level data integration to improve explainability in oncology AI systems. This structure ensures semantic alignment with biomedical ontologies like the National Cancer Institute Thesaurus (NCIt) and compliance with FAIR data principles, supporting interoperability and reusability for future biomedical graph research.

### B. Connectivity

The final graph comprises 1,340 nodes and 1,685 edges. A single **giant component** connects over 90% of nodes, showing strong biological integration. Diseases act as bridges between molecular (genes) and clinical (symptoms/treatments) layers.

### C. Community Detection

Using Leiden and Louvain algorithms, twelve communities were identified, each corresponding to major biomedical themes (TABLE I).

### D. Centrality Analysis

- **Degree**: High-degree nodes include Breast, Lung, and Leukemia.
- **Betweenness**: TP53 and Chemotherapy serve as inter-cluster bridges.
- **Eigenvector**: TP53 and BRCA1 remain globally influential.

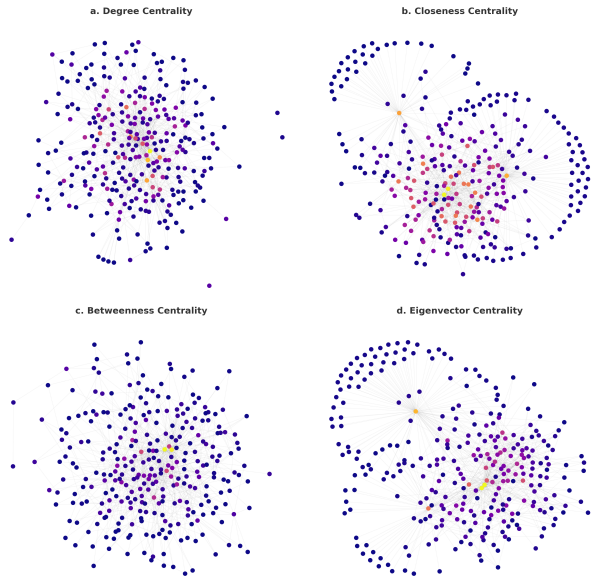


Fig. 2: Centrality metrics highlighting major hubs.

### E. Disease–Gene Subgraph

A subgraph focusing on diseases and their associated genes demonstrates the strongest biomedical coherence.

TABLE I: Cancer Clusters: Major Themes and Key Entities

| Cluster | Major Theme | Key Entities / Hubs |
|---|---|---|
| 1 | Breast / Ovarian / Endometrial cancers | **BRCA1**, **BRCA2**, estrogen, hormonal therapy |
| 2 | Lung / Esophageal / Colorectal cancers | **TP53**, **KRAS**, chemotherapy, smoking risk |
| 3 | Hematologic malignancies | **BCR-ABL**, **MYC**, leukemia, lymphoma, myeloma |
| 4 | CNS tumors (glioblastoma, medulloblastoma) | **EGFR**, **IDH1**, radiotherapy, temozolomide |
| 5 | Skin / Endocrine cancers (melanoma, thyroid) | **BRAF**, **RET**, MAPK pathway |
| 6 | Pediatric cancers (neuroblastoma, Wilms, sarcoma) | **ALK**, **RB1**, surgery, targeted therapy |
| 7–12 | Smaller clusters (biliary, prostate, head-and-neck, pancreatic, hepatic, rare) | Pathway-specific genes and localized therapies |

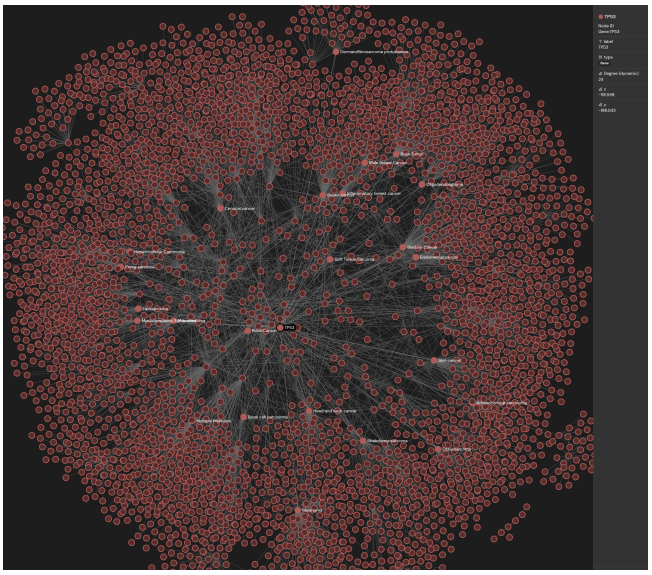

Fig. 3: Disease–gene subgraph showing genetic interconnectivity among cancers.

### F. Link Prediction

TABLE II: Predicted New Relationships

| Edge | Rationale |
|---|---|
| **Pancreatic** ↔ **BRCA2** | BRCA-driven DNA-repair defects extend to pancreatic tumors |
| **Ovarian** ↔ **TP53** | Shared apoptotic pathway alterations |
| **Endometrial** ↔ **Hormone therapy** | Consistent with estrogen-receptor modulation |
| **Glioblastoma** ↔ **EGFR inhibitors** | Pathway enrichment but underlinked in current data |

### G. Node Property Prediction

Neighbor-majority voting achieved approximately 93% accuracy in recovering node types, validating schema consistency.

### H. Traversal and Search

BFS from *Breast cancer*:

*Breast cancer* → **BRCA1** →

*DNA repair* → *Chemotherapy response*

DFS from *Colorectal cancer* reveals longer disease–gene–therapy chains, illustrating multi-hop reasoning.

## IV. Discussion

The analyses demonstrate a biologically faithful topology. Genes like TP53 and BRCA2 act as universal integrators, while Chemotherapy creates cross-domain connectivity. Community boundaries align with medical taxonomies, validating the extraction process. The graph structure reveals phenotype–genotype–therapy triads, where treatments cluster by mechanism (e.g., DNA damage response).

### A. Limitations

- Reliance on secondary textual sources limits molecular granularity.
- LLM extraction errors and deviations can introduce noise and limit reproducibility.
- Token constraints restrict final LLM analysis to a maximum of 300 diseases without summarization.

## V. Related Work

Prior biomedical KGs such as **Hetionet** and **BioKG** integrate curated databases (DrugBank, DisGeNET). Our system differs by extracting from public textual sources using LLMs, allowing automatic scaling. Menche et al. (Science, 2015) explored curated interactomes; our work complements this through automated extraction and modular graph analysis.

## VI. Conclusion

The cancer knowledge graph demonstrates high connectivity, biologically coherent modularity, and realistic centrality distributions.

- TP53 functions as a master hub across the oncological landscape.
- BRCA1/2 anchor hereditary cancer subnetworks.
- Chemotherapy connects molecular and clinical layers.

Future work includes edge weighting, integration of non-cancer diseases, and deployment of Graph Neural Networks (GraphSAGE, GAT) for predictive tasks.


## Acknowledgments

This work utilized the following LLM tools: ChatGPT, Google AI Studio. We would like to thank Djedje Didier GOHOUROU and Ruck THAWONMAS for their guidance and support.

## References


[1] National Library of Medicine. *MedlinePlus Web Service Developer Guide*, 2025.

[2] Wikipedia API Documentation. *MediaWiki Extracts Endpoint*, 2024.

[3] NetworkX Developers. *NetworkX: Complex Network Analysis in Python*, 2025.

[4] J. Menche et al., "Uncovering Disease–Disease Relationships through the Human Interactome," *Science*, vol. 347, no. 6224, 2015.

[5] Google AI Studio. *Gemini 2.5 Flash Live: Entity Extraction API Reference*, 2025.