# A ChatGPT Empowered Application in Biomedical Research

Judy Bai
Greenhills School, Ann Arbor, Michigan, USA

Logo created by researcher

All images, graphs, and diagrams developed by the researcher unless otherwise noted.

## Introduction

Accurately identifying the relationship between regulators and diseases has long been a challenge in the field of biomedical research due to its crucial roles in pathobiological conditions. MicroRNAs (miRNAs) play a crucial role in regulating gene expression and are implicated in a diverse range of human diseases. However, understanding their precise causal pathways remains challenging, primarily due to dispersed data annotation across databases. This underscores the urgent need for a unified data visualization format to streamline these resources, aiding the identification of biomarkers and therapeutic targets. The existing association databases rely on labor-intensive manual curation, hindering the timely addition of new associations from publications. This study addresses these challenges by developing the first unified miRNA-gene-disease knowledge graph, integrating ChatGPT API to automate miRNA-disease relation extraction from publications. The resulting knowledge graph not only enhances our understanding of miRNA involvement in human diseases but also provides a valuable resource for identifying biomarkers and therapeutic targets for future research.

## Background

**MicroRNA (MiRNA) and TarBase**

- MicroRNAs (miRNAs) regulate gene expression and are linked to human diseases (F1), but scattered data in databases complicates the understanding miRNA-gene-disease interactions.
- TarBase is an experimentally validated database consisting of miRNA and gene interactions through DIANA-microT.
- Consolidating fragmented data into a cohesive format is essential for facilitating the discovery of biomarkers and therapeutic targets related to miRNAs.
- Efforts are needed to develop a unified miRNA-gene-disease knowledge graph to enhance the understanding of causal pathways for potential clinical applications.

**ChatGPT and PubMed**

- ChatGPT is a large language model developed by OpenAI® and has been very powerful in natural language processing tasks to generate outputs based on user inputs (prompts).
- PubMed®, hosted by National Library of Medicine, includes over 36 million biomedical literature citations from MEDLINE, life science journals, and online books.
- Designing a meaningful prompt to interact with ChatGPT, especially a very accurate one to generate result files largely overlapping with ground-truth results, is a challenging task.
- Building a knowledge-based network graph by assembling multi-omics datasets is a crucial research area that can expedite the discovery of candidate disease biomarkers.

**Challenges of Fragmented Data**

- Fragmented data in databases hinders the understanding of miRNA causal pathways in gene regulation and disease development.
- Limited data cohesion affects research efficacy in identifying biomarkers and therapeutic targets related to miRNA .
- The creation of a unified miRNA-gene-disease knowledge graph aims to resolve data fragmentation challenges for improved research outcomes.

## Methodology

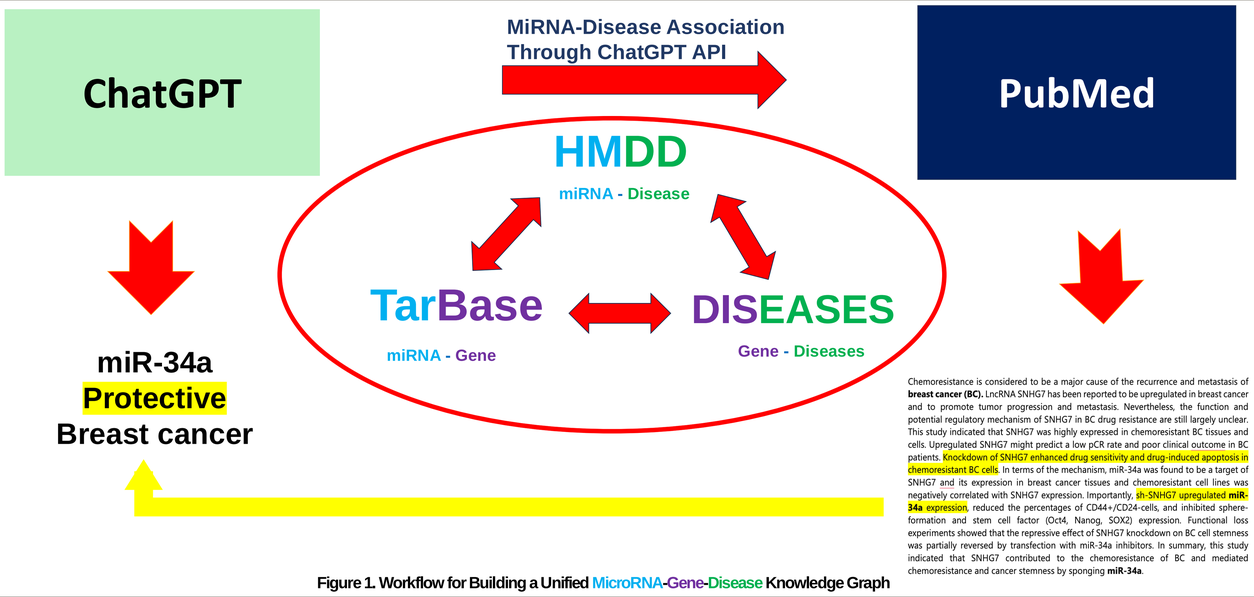
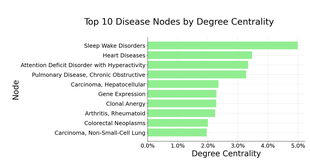



Figure 1. Workflow for Building a Unified MicroRNA-Gene-Disease Knowledge Graph

Chemoresistance is considered to be a major cause of the recurrence and metastasis of **breast cancer (BC).** LncRNA SNHG7 has been reported to be upregulated in breast cancer and to promote tumor progression and metastasis. Nevertheless, the function and potential regulatory mechanism of SNHG7 in BC drug resistance are still largely unclear. This study indicated that SNHG7 was highly expressed in chemoresistant BC tissues and cells. Upregulated SNHG7 might predict a low pCR rate and poor clinical outcome in BC patients. Knockdown of SNHG7 enhanced drug sensitivity and drug-induced apoptosis in chemoresistant BC cells. In terms of the mechanism, miR-34a was found to be a target of SNHG7 and its expression in breast cancer tissues and chemoresistant cell lines was negatively correlated with SNHG7 expression. Importantly, sh-SNHG7 upregulated **miR-34a** expression, reduced the percentages of CD44+/CD24-cells, and inhibited sphere-formation and stem cell factor (Oct4, Nanog, SOX2) expression. Functional loss experiments showed that the repressive effect of SNHG7 knockdown on BC cell stemness was partially reversed by transfection with miR-34a inhibitors. In summary, this study indicated that SNHG7 contributed to the chemoresistance of BC and mediated chemoresistance and cancer stemness by sponging **miR-34a**.

## Data Analysis and Results

- **Total number of entities: 34,492**
- **Total number of edges: 965,150**
- **Average Degree: 55.96**



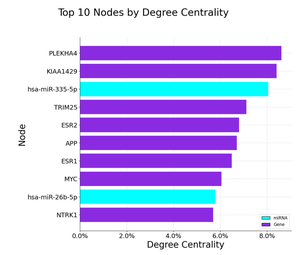
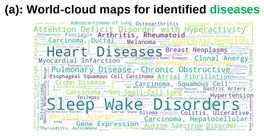




Figure 3: Top 10 nodes identified for disease, miRNA, and gene

(a): World-cloud maps for identified diseases




(b): World-cloud maps for identified miRNAs


(c): World-cloud maps for identified genes


Figure 4: Predicted world-cloud maps for identified (a) diseases (b) miRNAs (c) genes

## Results (Cont.)

- **ChatGPT API can provide high-accurate miRNA-disease relationship extractions from publication abstracts.**

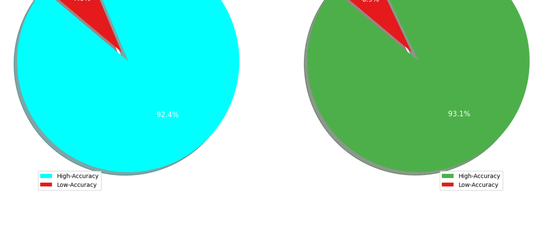






Figure 2: Accuracy of miRNA-diseases relationships extracted from publication abstracts

- Given two sets X and Y, the Jaccard coefficient can be computed as:

$$J(X,Y) = \left| \frac{X \cap Y}{X \cup Y} \right|$$

- where:
- $|X \cap Y|$ is the number of elements common to both sets X and Y,
- $|X \cup Y|$ is the total number of elements in either set X or Y (or both),
- J(X, Y) is the Jaccard coefficient, which is a number between 0 and 1.

Using Jaccard coefficient, I found that *IL6* (p<0.0001) can be associated with COVID-19 through their common shared miRNAs: *hsa-miR-29b-3p,hsa-miR-223-3p,hsa-miR-298,hsa-miR-142-3p,hsa-miR-98-5p,hsa-miR-451a,hsa-miR-19a-3p,hsa-miR-26a-5p,hsa-miR-125b-5p,hsa-miR-146a-3p,hsa-miR-146a-5p,hsa-miR-124-3p,hsa-miR-1-3p,hsa-miR-155-5p*

**Biopython NetworkX Library**

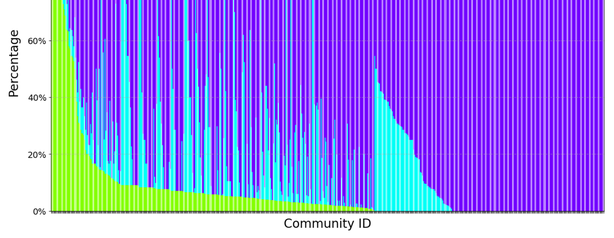



Figure 5: Communities detected by NetworkX with distributions of genes, miRNAs, and diseases

**Community Detection Results**

- **Most communities have the highest percentages of genes.**
- **Followed by gene-miRNA interactions, as represented by a large number of communities with a high percentage of miRNA and gene.**
- **This result demonstrated the fact that the most abundant data type currently in the field is miRNA-gene associations and gene-gene interactions, while gene-disease and miRNA-disease associations lag far behind.**

## Discussion and Conclusions

**Table 1. Enriched diseases identified by the knowledge graph**

| Enriched Disease | Expected | Observed | P-value | P-adjusted |
|---|---|---|---|---|
| **Acute lung injury** | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Aids dementia complex | 0.405162 | 11 | 1.90E-15 | 1.33E-13 |
| Antiphospholipid syndrome | 0.268283 | 10 | 2.40E-15 | 1.33E-13 |
| Atrophy | 0.268283 | 10 | 2.40E-15 | 1.33E-13 |
| Autoimmune diseases | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Cerebellar neoplasms | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Chemical and drug induced liver injury | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Germ cell and embryonal cancer | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Localized scleroderma | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Meningioma | 0.799374 | 13 | 5.43E-16 | 1.33E-13 |
| Myopia | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Odontogenic tumors | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Patau syndrome | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Radiation injuries | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| **Sars virus** | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |
| Vascular calcification | 0.273758 | 10 | 3.00E-15 | 1.33E-13 |

These microRNAs are enriched in various diseases, it shows that these miRNAs indeed are important regulators with diseases related to lung damage and corona virus infection.

**My knowledge graph can predict unknown gene-disease association**

- The unified miRNA-gene-disease knowledge graph created with my streamlines data extraction method, aiding in the identification of biomarkers and therapeutic targets for precision medicine.
- This approach represents a significant advancement in biomedical research by overcoming data fragmentation and enabling a more comprehensive understanding of miRNA-mediated disease pathways.
- The study's innovative use of ChatGPT API automates the extraction of miRNA-disease relationships from publications, contributing to improved disease management and personalized treatment strategies.

## Knowledge Graph

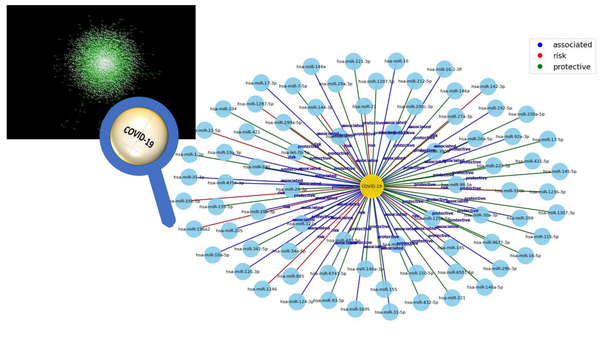



Figure 6. CASE STUDY: COVID-19. Next, I adopted two approaches to study the potential associations with COVID-19.

## CASE STUDY: COVID-19

COVID-19 is directly associated with miRNAs in the current literature, and no experimental evidence has been reported for gene-COVID-19 associations and COVID-19 comorbidity (although some literatures have investigated this before).

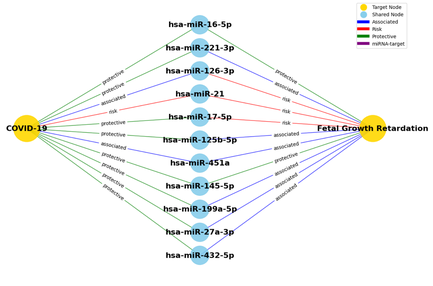



Figure 7. An association extraction to investigate COVID-19 comorbidity. All diseases within 2 steps away from COVID-19 were identified. The disease with the most shared miRNAs with COVID-19 is '**Fetal Growth Retardation**'.

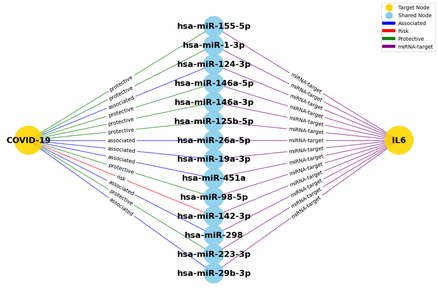



Figure 8. Novel association prediction to investigate gene and COVID-19 associations

## Selected References


- Wojciechowska A, Braniewska A, Kozar-Kamińska K. MicroRNA in cardiovascular biology and disease. Adv Clin Exp Med. 2017;26(5):865-874. doi:10.17219/acem/62915
- Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review [published correction appears in EMBO Mol Med. 2017 Jun;9(6):852]. EMBO Mol Med. 2012;4(3):143-159. doi:10.1002/emmm.201100209
- OpenAI. ChatGPT (Feb 24 version) [Large language model]. 2024; https://chat.openai.com.
- PubMed. United States: National Library of Medicine and National Center for Biotechnology Information; 2024. https://pubmed.ncbi.nlm.nih.gov/. Accessed February 25, 2024.
- Cui C, Zhong B, Fan R, Cui Q. HMDD v4.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 2024;52(D1):D1327-D1332. doi:10.1093/nar/gkad717
- Grissa D, Junge A, Oprea TI, Jensen LJ. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. Database (Oxford). 2022;2022:baac019. doi:10.1093/database/baac019.
- Huang HY, Lin YC, Li J, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. 2020;48(D1):D148-D154. doi:10.1093/nar/gkz896.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech.2008;P10008