BioNexus: A Global Network of Biomedical Relationships

Aaron Lee[1], **Benjamin Martinez**[1],

[1]Department of Computer Science, Stanford University

Correspondence: {aaroncl, benjm}@stanford.edu

Abstract

Extracting meaningful relationships between biomedical entities from unstructured text is a cornerstone for advancing research and clinical decision-making. Manual curation, however, is both time-intensive and resource-intensive, underscoring the need for scalable, automated solutions. This paper introduces an end-to-end system leveraging large language models (LLMs) for named entity recognition (NER) and relationship extraction (RE), designed to construct a prototype knowledge graph for biomedical research.

Our system processes PubMed abstracts to identify and normalize biomedical entities, extract relationships, and annotate findings with their sources metadata. Unlike traditional sentence-level methods, our paragraph-level approach enhances entity normalization and disambiguation by leveraging the contextual reasoning capabilities of LLMs. This enables the system to address common challenges, such as varying entity synonyms for the same entity and varying entity specificity within biomedical texts.

We produce a prototype knowledge graph with 170 PubMed abstracts demonstrating the effectiveness of the approach in creating structured knowledge graphs from unstructured data. We observe that our system can track paragraph level dependencies yet disambiguating entities remains a challenge. Future work will explore scaling domain-specific models and databases to improve the reliability and scope of the resulting knowledge graphs.

1 Introduction

Advancing biomedical research and clinical decision-making hinges on our ability to extract meaningful relationships between entities such as drugs, diseases, proteins, and genes from unstructured textual data. However, manual curation of these relationships is increasingly unsustainable due to the sheer volume of biomedical literature, with PubMed alone hosting millions of abstracts. This challenge highlights the critical need for automated methods capable of extracting, normalizing, and organizing biomedical knowledge into accessible formats.

The Global Network of Biomedical Relationships (GNBR) was a pivotal step in addressing this challenge (Percha and Altman, 2018). GNBR utilized sentence-level dependency analysis to process PubMed abstracts, identifying drug-disease interactions and generating structured knowledge graphs that supported predictive modeling and biomedical discovery. While effective, GNBR's reliance on traditional natural language processing (NLP) methods limited its ability to capture complex relationships and resolve ambiguities across varying contexts within the text. In general, other tools suffer from several limitations such as only annotating a small number of biomedical entity types. Second, they often use multiple single-type NER models to annotate entities of different types which limits generalizability. Additionally many tools employ models based on pre-defined rules with dictionaries (Wei et al., 2019), but they cannot cover complex variations of biomedical entity mentions. For instance, a simple dictionary-matching model cannot normalize 'oxichlorochine' into its canonical mention 'hydroxychloroquine' (mesh: D006886) unless the dictionary explicitly contains the mention 'oxichlorochine'.

In this work, we present GNBR 2.0, an evolution of the original GNBR framework that leverages the advanced contextual reasoning capabilities of large language models (LLMs). Through multiple LLM prompts, GNBR 2.0 introduces an end-to-end pipeline that automates named entity recognition (NER), relationship extraction (RE), and evidence aggregation from biomedical abstracts, producing a more nuanced and comprehensive knowledge graph. By expanding analysis from the sentence level to the paragraph level, GNBR 2.0 addresses key limitations of its predecessor, such as the disambiguation of hierarchical and overlapping entities and the normalization of terms with varying specificity.

Our primary contributions include:

1. Enhanced Named Entity and Relationship Extraction:

GNBR 2.0 performs multi-type NER and RE, identifying and contextualizing relationships between biomedical entities with improved precision.

2. Improved Entity Normalization and Disambiguation:

Using LLMs, our system resolves ambiguities by recognizing canonical names, aliases, and hierarchical distinctions, which are challenging for traditional methods.

3. Prototype Knowledge Graph Construction:

GNBR 2.0 produces evidence-rich, standardized knowledge graphs that include confidence scoring, metadata aggregation, and explicit references to supporting literature.

Through a proof-of-concept application to 170 PubMed abstracts, GNBR 2.0 demonstrates the potential of LLM-driven systems to create scalable, evidence-based knowledge graphs.

2 Related Work

Significant progress has been made in automating the extraction and understanding of biomedical relationships through various tools and frameworks. Embedding models, such as PubMedBERT, have played a transformative role by creating contextual representations fine-tuned for biomedical literature (Gu et al., 2021). While these models excel at recognizing domain-specific terminology, they tend to overfit their training data and often struggle with generalizing to real-world cases, particularly in tasks requiring nuanced relationship extraction.

Efforts to address these limitations have led to the development of advanced tools like AIONER (Luo et al., 2023). AIONER introduces an all-in-one (AIO) schema that integrates external annotated data to improve the accuracy, stability, and robustness of biomedical named entity recognition (BioNER). Its ability to recognize multiple entity types simultaneously and generalize to unseen entities marks a significant improvement over earlier methods. However, AIONER focuses exclusively on entity recognition and does not extend to the more complex task of relationship extraction and contextual analysis required for creating comprehensive biomedical knowledge graphs.

The BioRED track at BioCreative VIII introduced a benchmark challenge for extracting and categorizing relationships between biomedical entities, including identifying their semantic type and novelty. This initiative pushed the boundaries of relationship extraction by incorporating multiple entity types normalized to database identifiers, but the results highlighted persistent challenges. Teams often combined embedding models like PubMedBERT with large language models (LLMs) to improve performance, but even the best-performing systems achieved modest F-scores, particularly in the more comprehensive tasks of end-to-end extraction and novelty identification (Islamaj et al., 2024).

Tools like Pubtator and Metathesaurus have historically provided essential resources for biomedical text mining (Wei et al., 2024). Pubtator facilitates scalable annotation by linking entities to database identifiers, while the Metathesaurus consolidates biomedical terms and their relationships into a unified framework. These resources have proven invaluable for standardizing terminology, but their focus on entity annotation rather than extraction of relationships leaves gaps in addressing the complexities of biomedical texts.

3 Core Ideas

3.1 LLM Prompting Pipeline

1. Automated Data Collection and Parsing

The pipeline begins with an automated data collection system that retrieves biomedical abstracts from PubMed within user-specified date ranges. This system parses metadata such as title, authors, journal, and publication date, storing the information in a structured JSON format. By automating this step, GNBR 2.0 ensures scalability and adaptability for processing large and continuously growing biomedical literature databases.

2. Entity Retrieval and Normalization

GNBR 2.0 make LLMs calls to perform named entity recognition (NER) by identifying biomedical concepts and terms within abstracts (see Appendix, Listing 1 for an example). To enhance specificity, the system uses paragraph-level dependencies to refine entity terms. For instance, generic terms like "region 1" are contextualized into "region 1 of xyz protein" by analyzing surrounding text. The system also utilizes SciSpacy and Pubtator to link entities to unique biomedical database identifiers, addressing ambiguities and ensuring standardization across diverse sources.

3. Relationship Extraction with Contextual Evidence

The pipeline uses LLMs to identify predefined relationships between entities, such as drug-disease interactions or protein-protein interactions. The system ensures relationships are grounded in cited evidence by incorporating entity types, contextual dependencies, and relevant abstract sentences. Each extracted relationship is assigned a confidence score based on the strength and clarity of the contextual evidence. This context-aware extraction mitigates errors that arise from relying solely on co-occurrence or syntactic patterns.

4. Deduplication and Knowledge Graph Construction

After extracting entities and relationships, GNBR 2.0 normalizes the data into a structured graph format. Using a combination of rule-based algorithms and LLMs, the system merges highly similar entities and relationships into single nodes and edges, reducing redundancy. The resulting knowledge graph includes standardized nodes (representing biomedical entities such as drugs, genes, and diseases) and edges (defining semantic relationships), supplemented with metadata, confidence scores, and aggregated evidence references. This deduplication process ensures the graph's coherence and usability for downstream applications like predictive modeling and hypothesis generation.

4 Experiments

We attempted several prompting pipelines which will be described below. Evaluation is primarily qualitative analysis by manually inspecting abstracts and LLM outputs and identifying failure cases and patterns which will be discussed. This is because we focused on producing a semantically coherent graph on a small set of articles first.

4.1 Graph Prototype

We ran the above pipeline on 170 with GPT-4o-mini on randomly scraped PubMed articles. The result graph includes nodes annotated with a standardized node name, aliases, type (protein, gene, drug, disease, chemical), description. The edges are annotated with source node, target node, citation of original paper, cited text of where the relation is taken from the abstract. For each relation there is also metadata such as aggregated data e.g. how many papers corroborate the relation, last updated, and a confidence score which is how much the LLM believes that the relation is accurate. The prototype is available here (see README instructions to visualize the graph).

As seen in Figure 1, we also ran the above pipeline on a singular abstract from the paper "Structural and functional insights into the major mutations of SARS-CoV-2 Spike RBD and its interaction with human ACE2 receptor" alongside a knowledge graph generated by our model. The abstract provides insight on the molecular mechanisms underlying the binding affinities of spike protein mutations (L452R, T478K, and N501Y) to the ACE2 receptor and their implications for SARS-CoV-2.

In the abstract, sentences extracted to define relationships are highlighted with edge colors corresponding to the relationships in the graph. The knowledge graph captures three key types of relationships: ASSOCIATE (between SARS-CoV-2 and L452R, T478K, N501Y), POSITIVE_CORRELATE (between L452R, T478K, N501Y and ACE2), and INTERACT (between Spike RBD and ACE2). Unlike earlier systems that failed to detect the relationships between all three proteins and COVID-19 since they do not cooccur in the same sentence, our model demonstrates the ability to integrate information across sentences throughout the entire abstract and accurately represent those relationships.

4.2 XML Tagging Entities

BioRED provides a human annotated dataset of abstracts as a gold standard (Luo et al., 2022). The entity annotations they provide are not normalized and their relation annotations have a custom set of rules to bucket a named relation into their standardized relation. Their entity annotations are in the form of XML tags i.e. if a term in the text is a biomedical entity then it will be surrounded by its type e.g. <DiseaseOrPhenotypicFeature>"long QT syndrome"</DiseaseOrPhenotypicFeature>. We wanted to compare how well our LLM entity recognizer can perform this task so we prompted the LLM, with a similar pipeline to the steps outlined in section 3, we produce these XML tags for 100 abstracts and evaluated with results in the table below. BioRED tested their own LLM prompting method on these 100 abstracts which serve as our baseline below (Islamaj et al., 2024). Future work can be done to explore this method and extend it to relationship extraction which we did not primarily focus on.

Method	Precision	Recall	F1
Ours	**0.40**	0.64	**0.49**
BioRed	0.28	**0.96**	0.42

Table 1: NER evaluated on BioRED's 100 hand-annotated dataset; BioRED provides a GPT baseline.

5 Observations and Insights

Improved Contextual Entity Normalization: GNBR 2.0 demonstrates notable improvements in addressing entity over-fetching and redundancy compared to prior approaches. Traditional methods, such as sentence-level analysis and embedding-based models like PubMedBERT, struggled with niche cases involving ambiguous or repetitive references to entities. While the initial LLM outputs mirrored these challenges, the iterative use of paragraph-level contextual analysis enabled the system to detect and refine ambiguities effectively. For instance, when encountering multiple references to regions of "apo B100," the system successfully normalized and grouped the terms into a coherent representation. This capability highlights the advantages of leveraging LLMs for nuanced entity normalization.

Hierarchical Disambiguation Remains a Challenge: Despite advancements, GNBR 2.0 faces challenges in distinguishing hierarchical relationships between entities. For example, the LLM often grouped distinct terms like "Crohn's disease" and "ulcerative colitis" under the broader category of "inflammatory bowel disease," obscuring the unique characteristics of the subcategories. This limitation underscores the difficulty of maintaining granularity in hierarchical distinctions, a problem that remains unresolved in the current iteration of the system.

Overproduction of Irrelevant Entities: Our design of the entity extractor encouraged liberally gathering any and all biomedical entities so that we have a comprehensive set. But this also resulted in irrelevant entities being produced that were overtly not biomedical. For instance, it could classify an entity as "Italy" with type "Gene". Better prompt engineering and other LLM calls to filter non-medical entities can be implemented to address this.

Underproduction of Relationship Diversity: The relationship extraction component tends to prioritize frequent and well-known associations, under-producing less common or novel relationships. While this behavior ensures reliability in extracting established relationships, it limits the system's ability to uncover novel associations that could drive scientific discovery. This shortfall suggests a need for improved mechanisms to identify rare but meaningful relationships in biomedical literature.

Canonical Naming and Database Integration: GNBR 2.0 successfully generates canonical names for entities, providing a standardized representation that facilitates downstream analysis. However, the system's scalability is hindered by its limited integration with database identifiers, such as those provided by Pubtator or the Metathesaurus.

Delete: These identifiers are essential for disambiguating entities at scale and ensuring consistency across diverse datasets. Addressing this gap will be critical for enhancing the precision and utility of the resulting knowledge graph.

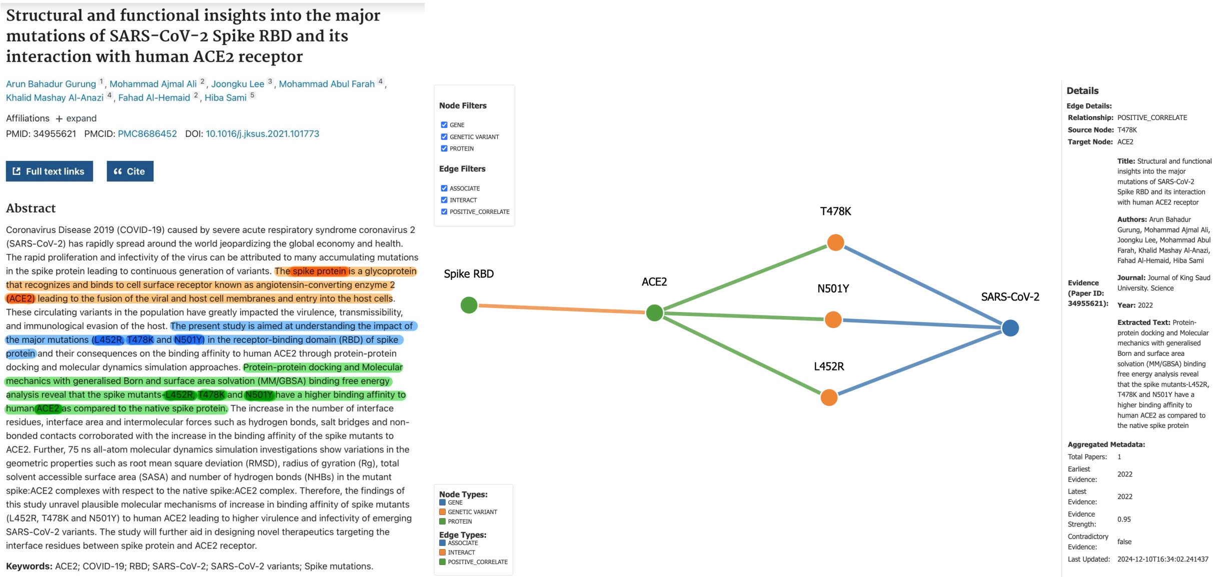
6 Lessons and Future work

6.1 Grounded Entity Disambiguation and Relation Identifiers

Entity disambiguation posed one of the largest challenges of the project since many biomedical entities carry different, and often similar, names and some biomedical terms may refer to multiple entities. These entities also present themselves in a myriad of different and ever-evolving contexts, making it incredibly difficult to broadly navigate all of biomedical nomenclature.

Our system's capability in normalizing and grounding entities in domain-specific databases is limited. We made attempts at constructing an ensemble system, but future work involves continuing to integrate and scale domain-specific NER models and tools such as All-In-One Named Entity Recognition (AIONER), SciSpacy and Pubtator 3.0. None of the attempts we made consistently achieved the accuracy and reliability required for a fully grounded biomedical knowledge graph.

Integrating this deterministic step into our pipeline would have seen some sort of fast lookup mechanism between the retrieved entity (by the first generic LLM call). In this case, we can provide the LLM with more context to extract entities incorporate data from multiple libraries and traditional en-



Structural and functional insights into the major mutations of SARS-CoV-2 Spike RBD and its interaction with human ACE2 receptor

Arun Bahadur Gurung [1], Mohammad Ajmal Ali [2], Joongku Lee [3], Mohammad Abul Farah [4], Khalid Mashay Al-Anazi [4], Fahad Al-Hemaid [2], Hiba Sami [5]

Affiliations + expand

PMID: 34955621 PMCID: PMC8686452 DOI: 10.1016/j.jksus.2021.101773

Full text links | Cite

Abstract

Coronavirus Disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has rapidly spread around the world jeopardizing the global economy and health. The rapid proliferation and infectivity of the virus can be attributed to many accumulating mutations in the spike protein leading to continuous generation of variants. The spike protein is a glycoprotein that recognizes and binds to cell surface receptor known as angiotensin-converting enzyme 2 (ACE2) leading to the fusion of the viral and host cell membranes and entry into the host cells. These circulating variants in the population have greatly impacted the virulence, transmissibility, and immunological evasion of the host. The present study is aimed at understanding the impact of the major mutations (L452R, T478K and N501Y) in the receptor-binding domain (RBD) of spike protein and their consequences on the binding affinity to human ACE2 through protein-protein docking and molecular dynamics simulation approaches. Protein-protein docking and Molecular mechanics with generalised Born and surface area solvation (MM/GBSA) binding free energy analysis reveal that the spike mutants-L452R, T478K and N501Y have a higher binding affinity to human ACE2 as compared to the native spike protein. The increase in the number of interface residues, interface area and intermolecular forces such as hydrogen bonds, salt bridges and non-bonded contacts corroborated with the increase in the binding affinity of the spike mutants to ACE2. Further, 75 ns all-atom molecular dynamics simulation investigations show variations in the geometric properties such as root mean square deviation (RMSD), radius of gyration (Rg), total solvent accessible surface area (SASA) and number of hydrogen bonds (NHBs) in the mutant spike:ACE2 complexes with respect to the native spike:ACE2 complex. Therefore, the findings of this study unravel plausible molecular mechanisms of increase in binding affinity of spike mutants (L452R, T478K and N501Y) to human ACE2 leading to higher virulence and infectivity of emerging SARS-CoV-2 variants. The study will further aid in designing novel therapeutics targeting the interface residues between spike protein and ACE2 receptor.

Keywords: ACE2; COVID-19; RBD; SARS-CoV-2; SARS-CoV-2 variants; Spike mutations.

Node Filters

☑ GENE
☑ GENETIC VARIANT
☑ PROTEIN

Edge Filters

☑ ASSOCIATE
☑ INTERACT
☑ POSITIVE_CORRELATE

Node Types:
🟦 GENE
🟧 GENETIC VARIANT
🟩 PROTEIN

Edge Types:
🟦 ASSOCIATE
🟧 INTERACT
🟩 POSITIVE_CORRELATE

Details

Edge Details:

Relationship:	POSITIVE_CORRELATE
Source Node:	T478K
Target Node:	ACE2
Evidence (Paper ID: 34955621):	**Title:** Structural and functional insights into the major mutations of SARS-CoV-2 Spike RBD and its interaction with human ACE2 receptor **Authors:** Arun Bahadur Gurung, Mohammad Ajmal Ali, Joongku Lee, Mohammad Abul Farah, Khalid Mashay Al-Anazi, Fahad Al-Hemaid, Hiba Sami **Journal:** Journal of King Saud University. Science **Year:** 2022 **Extracted Text:** Protein-protein docking and Molecular mechanics with generalised Born and surface area solvation (MM/GBSA) binding free energy analysis reveal that the spike mutants-L452R, T478K and N501Y have a higher binding affinity to human ACE2 as compared to the native spike protein

Aggregated Metadata:

Total Papers:	1
Earliest Evidence:	2022
Latest Evidence:	2022
Evidence Strength:	0.95
Contradictory Evidence:	false
Last Updated:	2024-12-10T16:34:02.241437

Figure 1: An isolated example on a difficult abstract, demonstrating the system's ability to infer relationships even when the disease name and variants do not cooccur within the same sentence. The sentences extracted are highlighted with edge colors corresponding to the relationships in the graph.

tity parsers to strengthen the reliability of the drawn relationships. One avenue we heavily researched was entity assignment of a concept unique identifier (CUI), which is used to uniquely identify concepts in the Unified Medical Language System (UMLS) Metathesaurus by the National Institutes of Health. However, while these domain-specific tools can offer structured mechanisms for entity disambiguation, their deterministic and rule-based approaches struggled when presented with incomplete contexts, ambiguous abbreviations, and terms without pre-existing stable identifiers. Instead, we could look towards a more tightly integrated, context-sensitive approach that would allow for iterative refinement between the LLM and domain-specific resources. In lieu of our linear 5-step pipeline (LLM first with deterministic lookup second or generic LLM calls throughout), a feedback loop could prove beneficial, where the LLM can request additional clarification or leverage intermediate IDs from SciSpacy or Pubtator to refine its entity normalizations and relationship detections.

6.2 Knowledge Graph Design for Complex Relations

Biomedicine is a particularly challenging domain for its minute differences between entities, variations in naming, and complex relations. When manually looking at abstracts we found certain difficult cases to encode into a knowledge graph. For example, "Vascular endothelial cells respond to TNF by undergoing a number of pro-inflammatory changes, which increase leukocyte adhesion." In other words, TNF causes the inflammation of these endothelial cells which increase leukocyte adhesion. One possibility to encode this is to name 3 entities: "TNF", "Inflammation of endothelial cells", "leuokocyte adhesion". However, the term 'Inflammation of endothelial cells' can be both ambiguous and not standard enough to normalize. To our knowledge, standard methods of modeling knowledge graphs do not handle such cases accurately as they only have enumerated node and edge types. One method to resolve such a case would be to introduce a new design to model knowledge graphs such that it captures hierarchical distinctions and more refined types. For example the encoding could designate certain conditions like 'inflamed' and relation types that modify an entity to transition into that condition. For example, the process becomes "TNF" triggers the condition "inflamed" onto targets ["endothelial cells",...] which produces "inflamed endothelial cells", so that we've modeled "inflamed endothelial cells" increase leukocyte adhesion while non-inflamed endothelial cells don't. There are other complex cases which should be considered. There is future work to be done on this topic of knowledge graph design for LLM enabled graph construction.

6.3 Handling Contradiction

Another avenue of research could analyze the best methods to contend with variance in biomedical literature. Many questions in biomedicine lack a general scientific consensus and new research continually upends previously established findings. Although we do already ask the LLM to asking a confidence score, we can look further into tracking the number of citations of a certain relationship that was discovered to make a more accurate calculation of confidence. Manually assessing the quality of the underlying studies and leveraging other meta-analyses and literature surveys could also provide a more robust framework for evaluating confidence.

7 Conclusion

In this work, we presented GNBR 2.0, an end-to-end system leveraging large language models to extract and organize biomedical knowledge from unstructured text into a prototype knowledge graph. By incorporating paragraph-level contextual reasoning, the system addresses limitations of traditional sentence-level methods, such as entity ambiguity and redundancy, and enhances the precision of relationship extraction. Through a proof-of-concept evaluation on 170 PubMed abstracts, GNBR 2.0 demonstrated its ability to normalize entities, extract evidence-grounded relationships, and aggregate metadata for structured knowledge representation. Despite these advancements, challenges remain in hierarchical disambiguation and capturing less frequent, novel relationships.

Looking forward, GNBR 2.0 provides a foundation for advancing automated biomedical knowledge extraction, highlighting the potential of LLM-driven approaches in addressing domain-specific challenges. Future work will focus on improving entity grounding with database identifiers, scaling to include comprehensive datasets, and refining knowledge graph design to handle complex relationships and hierarchical distinctions. These efforts aim to enhance the reliability, scalability, and utility of biomedical knowledge graphs, empowering research and clinical decision-making through more robust and comprehensive data representations.

References

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard A. A Jonker, Sofia I. R Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, Jiru Li, Dinghao Pan, Wilailack Meesawad, Richard Tzong-Han Tsai, M. Janina Sarol, Gibong Hong, Airat Valiev, Elena Tutubalina, Shao-Man Lee, Yi-Yu Hsu, Mingjie Li, Karin Verspoor, and Zhiyong Lu. 2024. The overview of the biored (biomedical relation extraction dataset) track at biocreative viii. *Database*, 2024:baae069.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

Bethany Percha and Russ B Altman. 2018. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15):2614–2624.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. 2024. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

A Example Appendix

```
Prompt:

You are an expert in biomedical
    information extraction. Your task is
     to process a PubMed biomedical
    abstract and identify named entities
    . Limit your extraction to the
    following entity types: Drug,
    Disease, Gene, and Protein.

For each entity identified, return a
    structured JSON object with the
    following fields:

    entity_name: The name of the entity,
         described as specifically as
        possible. If the entity name is
        not descriptive (e.g., "Region
        1"), enrich it using contextual
        information from the abstract.
        For example:
```

```
        If the text mentions "Region 1
            (0014Lys-Ser0160)" and
            relates it to "Apo B100,"
            return the name as "Region 1
             of Apo B100 (0014Lys-
            Ser0160)."
    type: The type of entity (Drug,
        Disease, Gene, or Protein).
    description: A short description of
        the entity, if available. This
        can be derived from the abstract
         text.
    is_ambiguous: A boolean field (true
        or false). Set this to true if
        the entity name cannot be made
        specific or if there isnt
        enough context to resolve
        ambiguity.

Rules:

    If an entity name is not inherently
        descriptive, enhance it using
        additional context from the
        abstract.
    If there isnt sufficient context
        to create a specific name, flag
        the entity as is_ambiguous: true
         for further processing.
    Ensure that ambiguous entities
        retain their original names but
        include the ambiguity flag.
    Exclude entities that dont fall
        into the specified types.

Input:

Here is the PubMed biomedical abstract
    for processing:

<Insert abstract here>

Example Output:

[
  {
    "entity_name": "Aspirin",
    "type": "Drug",
    "description": "A drug commonly used
         for pain relief and anti-
        inflammatory purposes.",
    "is_ambiguous": false
  },
  {
    "entity_name": "Region 1 (0014Lys-
        Ser0160)",
    "type": "Protein",
    "description": "A region identified
        in Apo B100.",
    "is_ambiguous": true
  },
  {
    "entity_name": "Alzheimer's disease
        ",
    "type": "Disease",
    "description": "A progressive
        neurodegenerative disorder that
        affects memory and cognitive
        function.",
    "is_ambiguous": false
  }
]
```

Listing 1: Example prompt to gather entities

```
You are an expert in biomedical
    information extraction. Your task is
     to analyze a given biomedical
    abstract and tag all entities in the
     text into specific categories. You
    will return the abstract with XML-
    like tags inserted around the
    identified entities.
Categories for Tagging:

    GeneOrGeneProduct: For genes,
        proteins, mRNA, and other gene
        products.
    ChemicalEntity: For chemicals and
        drugs.
    DiseaseOrPhenotypicFeature: For
        diseases, symptoms, and disease-
        related phenotypes.
    SequenceVariant: For genomic or
        protein variants, including
        substitutions, deletions,
        insertions, and others.
    OrganismTaxon: For species in the
        hierarchical taxonomy of
        organisms.
    CellLine: For cell lines.

Rules:

    Comprehensive and Liberal Tagging:
        Be inclusive and identify all
            possible entities within the
             abstract.
        Add tags even if the context is
            ambiguous, but ensure the
            tags are consistent with the
             entity type.

    Tag Format:
        Use the format <info type="
            Category">text</info> where
            Category is one of the
            predefined entity types (e.g
            ., GeneOrGeneProduct,
            ChemicalEntity).
        Example: <info type="
            GeneOrGeneProduct">BRCA1</
            info> for a gene.

    Preserve Abstract Structure:
        Maintain the original text
            structure and spacing of the
             abstract.
        Insert tags seamlessly around
            the identified entities.

    Handle Ambiguous Cases:
        If an entity could belong to
            multiple categories, choose
            the most specific category
            based on context.

Input Example:

Abstract:
```

```
Aconitine binds to Na(+) channels,
    leading to arrhythmia and
    cardiotoxicity. It also increases
    intracellular Ca2+, enhancing
    myocardial injury through the
    activation of p38 MAPK pathways.

Output Example:

<info type=”ChemicalEntity”>Aconitine</
    info> binds to <info type=”
    GeneOrGeneProduct”>Na(+)</info>
    channels, leading to <info type=”
    DiseaseOrPhenotypicFeature”>
    arrhythmia</info> and <info type=”
    DiseaseOrPhenotypicFeature”>
    cardiotoxicity</info>. It also
    increases intracellular <info type=”
    ChemicalEntity”>Ca2+</info>,
    enhancing <info type=”
    DiseaseOrPhenotypicFeature”>
    myocardial injury</info> through the
     activation of <info type=”
    GeneOrGeneProduct”>p38 MAPK</info>
    pathways.

Notes:

    Entity Context:
        Include any modifiers or
            surrounding text that
            clarify the entity (e.g., ”
            human BRCA1 gene” should be
            fully tagged as <info type=”
            GeneOrGeneProduct”>human
            BRCA1 gene</info>).

    Nested Entities:
        Avoid overlapping or nested tags
            ; each entity should be
            tagged independently.

    Be Exhaustive:
        Err on the side of over-tagging
            rather than under-tagging,
            ensuring the abstract is
            thoroughly annotated.

    Return Only the Tagged Abstract:
        Do not include any additional
            comments or explanations in
            the output.
```

Listing 2: Example prompt to tag entities