Journal of AI, ML, and Bioinformatics 2025, pp. 47–54
https://doi.org/10.5147/jaimlb.267

Published in the United States of America
https://www.atlas-publishing.org

Mining Alzheimer's–Related Gene Mentions from PubMed Using NLP and Enrichment Analysis: A Temporal and Network Perspective

My Abdelmajid Kassem[1*], Khalid Lodhi[1], Youssef Jouad[2], and Jiazheng Yuan[1*]

[1] Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA; [2] IT Programs Data Center, Durham Technical Community College, Durham, NC 27703, USA.

Received: December 31, 2025 / Accepted: February 8, 2026

Abstract

Alzheimer's disease (AD) remains a leading cause of morbidity and mortality worldwide, with genetics playing a critical role in disease onset and progression. However, systematically mapping the evolving landscape of gene-focused AD research remains challenging due to the rapid growth of biomedical literature. We applied large-scale named entity recognition using BioBERT NER on 9,742 PubMed abstracts from 2010 to 2023 related to AD genetics. Entities were extracted, quantified, and visualized using bar plots, temporal trend analyses, and co-mention network graphs. Enrichment analysis was performed using the Enrichr API on top-mentioned gene entities. "AD" and "Alzheimer" dominated mentions across the dataset, validating the retrieval strategy. Geographical trends aligned with global research output, while co-mention networks revealed thematic clustering between AD, Alzheimer's disease, and key genes. Temporal trends demonstrated consistent focus on top genes over 14 years, underscoring stable scientific interest in genetic underpinnings of AD. Enrichment analysis confirmed associations with known neurodegenerative pathways. This study highlights the feasibility and value of scalable biomedical NER and network analysis to map and monitor the research landscape of AD genetics. The workflow provides a quantitative foundation for tracking emerging gene targets and research gaps, facilitating hypothesis generation and informed prioritization in neurodegenerative research.

Keywords: Alzheimer's disease, Gene mining, Named entity recognition, BioBERT, PubMed text mining, Co-mention networks, Temporal trends, Biomedical informatics, Neurodegenerative diseases, Literature surveillance.

* Corresponding author: mkassem@uncfsu.edu

Journal of Artificial Intelligence, Machine Learning, and Bioinformatics - ISSN 3068-5036. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia, characterized by progressive memory loss, cognitive impairment, and functional decline, posing a substantial global health burden (Scheltens et al., 2021). With an estimated 50 million people affected worldwide, and projections indicating a threefold increase by 2050, there is an urgent need to deepen our understanding of the molecular mechanisms underlying AD to enable early diagnosis, effective therapeutic interventions, and prevention strategies (Livingston et al., 2020; Gaugler et al., 2023).

The development of Alzheimer's disease is driven by a multifaceted interaction among genetic predispositions, environmental exposures, and lifestyle influences. At the molecular level, hallmark features include extracellular amyloid-beta (Aβ) plaque deposition and intracellular neurofibrillary tangles composed of hyperphosphorylated tau protein, which lead to synaptic dysfunction and neuronal loss (Long and Holtzman, 2019). Among the genetic risk factors, mutations in APP (Hardy and Selkoe, 2002), PSEN1 (De Strooper, 2007), and PSEN2 (Rogaeva et al., 2001) (Table 1) are causative in familial early-onset AD, while the APOE ε4 allele (Corder et al., 1993) (Table 1) is the strongest genetic risk factor for sporadic late-onset AD (Tanzi, 2012; Belloy et al., 2019). Despite these known contributors, the genetic architecture of AD remains incompletely understood, with ongoing genome-wide association studies (GWAS) and transcriptomic analyses continuously identifying novel susceptibility loci and potential therapeutic targets (Kunkle et al., 2019; Andrews et al., 2023).

The biomedical literature on AD is expanding rapidly, with thousands of articles published annually on the disease's genetic underpinnings, mechanisms, and potential interventions (Wozniak et al., 2024). Manually curating, synthesizing, and extracting insights from this literature is infeasible, necessitating computational methods for scalable literature mining (Wang et al., 2022). Natural Language Processing (NLP) has emerged as a powerful tool in biomedical informatics, enabling automated extraction of entities (e.g., genes, proteins, diseases) and relationships from unstructured text, facilitating systematic reviews, hypothesis generation, and knowledge discovery (Wei et al., 2019; Lee et al., 2020; Wang et al., 2022; Zhang et al., 2024; Shakeri et al., 2025).

Recent advances in NLP, particularly the advent of deep learning models such as transformer-based architectures, have significantly improved the accuracy and scalability of entity recognition and relation extraction in biomedical texts (Madan et al., 2024; Mandal et al., 2025). Models like PubTator Central (Wei et al., 2019), BERT (Alsentzer et al., 2019), BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019) have demonstrated strong performance in biomedical named entity recognition (NER), question answering, and document classification tasks by leveraging contextual embeddings from large-scale biomedical corpora. These tools enable high-throughput analysis of PubMed abstracts, allowing researchers to track research trends, identify emerging biomarkers, and map complex disease-related molecular networks (Ogunjobi et al., 2024).

Literature mining has been applied to AD to identify gene-disease associations, highlight underexplored molecular targets, and support drug repurposing efforts (Singh et al., 2022; Sikirzhytskaya et al., 2024). For example, Liu et al. (2024) leveraged text mining to extract AD-related genes from literature and combined these with GWAS data to identify novel candidate genes for experimental validation. Similarly, Shakeri et al. (2025) applied NLP approaches to extract risk factors from AD-related publications to build risk models, demonstrating the potential of NLP to accelerate translational research in neurodegenerative diseases.

In addition to entity extraction, coupling NLP-derived gene lists with gene set enrichment analysis allows the contextualization of findings within biological pathways and functional categories, providing a system-level perspective on the literature landscape (Kuleshov et al., 2016). Pathway enrichment using tools like Enrichr enables the identification of significantly overrepresented pathways, such as amyloid fiber formation, synaptic signaling, and neuroinflammatory processes, which are central to AD pathology (Yuen et al., 2020). Furthermore, temporal trend analysis of gene mentions across years provides insight into evolving research interests and emerging molecular targets, supporting funding prioritization and systematic reviews (Neveol et al., 2018).

Network analysis, such as gene co-mention networks derived from literature, offers additional layers of insight by highlighting frequently co-occurring gene pairs, suggesting potential interactions or shared involvement in biological processes (Junge and Jensen, 2020; Shahri et al., 2021). Such networks can reveal central hub genes, which may play pivotal roles in disease mechanisms and represent promising therapeutic targets.

Despite these advancements, there is a need for a scalable, end-to-end pipeline that integrates transformer-based NLP for entity extraction, pathway enrichment, temporal trend analysis, and network visualization in a single framework, specifically tailored to AD literature mining. This integrated approach will enable researchers to systematically track emerging genes, identify significant pathways, and visualize gene interaction patterns, ultimately supporting hypothesis generation and targeted experimental validation in AD research.

In this study, we present a scalable pipeline leveraging transformer-based NLP models, PubMed data mining, gene set enrichment analysis, and network visualization to systematically extract and analyze gene mentions related to Alzheimer's disease from 2010 to 2023. We applied named entity recognition to over 9,700 abstracts, tracked temporal trends of top genes, constructed gene co-mention networks, and performed KEGG pathway enrichment analysis on top gene mentions. This pipeline aims to support researchers in identifying key genetic contributors, tracking evolving research focuses, and contextualizing findings within biological pathways to advance understanding and intervention strategies for Alzheimer's disease.

2. Methods

2.1. Data Collection from PubMed

We systematically retrieved abstracts from PubMed to capture the breadth of Alzheimer's disease (AD) genetic research from 2010 to 2023 using the NCBI Entrez API (Sayers et al., 2024). The search query used was "Alzheimer AND genes," ensuring inclusion of studies focusing on the genetic underpinnings of AD across molecular, clinical, and translational research contexts. For each year, up to 1,000 abstracts were fetched depending on availability, with a one-second delay between requests to comply with NCBI API guidelines. Abstracts were retrieved in plain text, parsed, and preprocessed to remove duplicates and empty entries, resulting in a curated corpus of 9,742 abstracts for downstream analysis.

Table 1. Alzheimer's disease related gene acronyms and definitions. This table summarizes the key genes and alleles frequently identified in Alzheimer's disease (AD) literature and extracted through named entity recognition (NER) during this study. Each entry includes the full gene name, known biological function, and its established or hypothesized role in AD pathogenesis.

Acronym	Full Name	Function and Role in Alzheimer's Disease	Reference
APP	*Amyloid Precursor Protein*	Encodes a membrane protein cleaved to produce β-amyloid (Aβ) peptides; Aβ aggregation is a central hallmark of Alzheimer's pathology.	Hardy and Selkoe, 2002
PSEN1	*Presenilin 1*	Part of the γ-secretase complex that cleaves APP. Mutations cause early-onset familial Alzheimer's by increasing Aβ42 production.	De Strooper, 2007
PSEN2	*Presenilin 2*	Homologous to PSEN1; also, part of the γ-secretase complex. Mutations are less common but associated with early-onset Alzheimer's.	Rogaeva et al., 2001
APOE ε4	*Apolipoprotein E epsilon 4 allele*	A genetic risk factor for late-onset Alzheimer's disease. The ε4 variant is associated with increased Aβ deposition and reduced clearance from the brain.	Corder et al., 1993

Journal of Artificial Intelligence, Machine Learning, and Bioinformatics - ISSN 3068-5036. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

Using PubMed abstracts as a data source has been validated in prior studies to effectively reflect biomedical research trends while maintaining scalability and accessibility (Table 2) (Neveol et al., 2018; Wallach et al., 2018).

2.2. Named Entity Recognition Using Transformer Models

For named entity recognition (NER), we leveraged transformer-based models via the Hugging Face Transformers library (Keraghel et al., 2024), which has demonstrated state-of-the-art performance in biomedical text mining tasks (Lee et al., 2020; Beltagy et al., 2019). Specifically, we utilized the dslim/bert-base-NER model, which is fine-tuned for general-purpose NER and provides robust extraction of person, organization, and location entities. we used the "PER" label as a proxy for gene and gene product mentions, acknowledging that while not exclusively capturing gene entities, this approach provides a practical and scalable approximation in the absence of domain-specific tokenization within the model (Zhang et al., 2024). The abstracts were batch-processed to extract entities and associated entity groups while retaining their positional and textual context for further aggregation and analysis (Table 2).

2.3. Entity Aggregation and Frequency Analysis

Extracted entities were aggregated and categorized based on their labels, focusing on "PER" to represent gene mentions, with mention counts accumulated across the corpus and stratified by publication year. This enabled identification of frequently mentioned genes and gene products in AD literature, facilitating prioritization for pathway analysis and temporal trend visualization. Similar entity extraction pipelines have successfully enabled scalable literature mining in biomedical contexts, allowing for the identification of high-impact genes and emergent molecular targets in diseases such as cancer and neurodegeneration (Table 2).

2.4. Temporal Trend Analysis

To investigate evolving research focuses, we conducted temporal trend analysis of gene mentions by calculating annual mention frequencies of the top identified gene entities from 2010 to 2023. This analysis allowed us to visualize the dynamics of gene-related research attention over time, identifying emerging genes of interest and persistent key players within AD research. Such temporal analyses have proven valuable in bibliometric studies, enabling the detection of shifting scientific priorities and emerging molecular hypotheses in biomedical research (Table 2) (Neveol et al., 2018; Zhang and Fan, 2022).

2.5. Gene Co-Mention Network Construction

To map potential associations between genes discussed in the AD literature, we constructed gene co-mention networks based on co-occurrence within the same abstract. For each abstract, we extracted the set of unique gene mentions and created undirected edges between all possible gene pairs within that abstract, incrementing edge weights with each additional co-mention occurrence across the corpus. We then filtered and visualized the top 50 edges with the highest weights to create interpretable networks highlighting hub genes and co-discussed gene clusters. Co-mention network analysis serves as a proxy for potential functional, experimental, or thematic relationships between genes and has been used effectively in literature mining studies to generate hypotheses for further experimental validation (Table 2) (Junge and Jensen, 2020; Liu et al., 2024).

2.6. Pathway Enrichment Analysis Using Enrichr

To contextualize the top mentioned genes within biological pathways, we performed pathway enrichment analysis using Enrichr (Kuleshov et al., 2016),

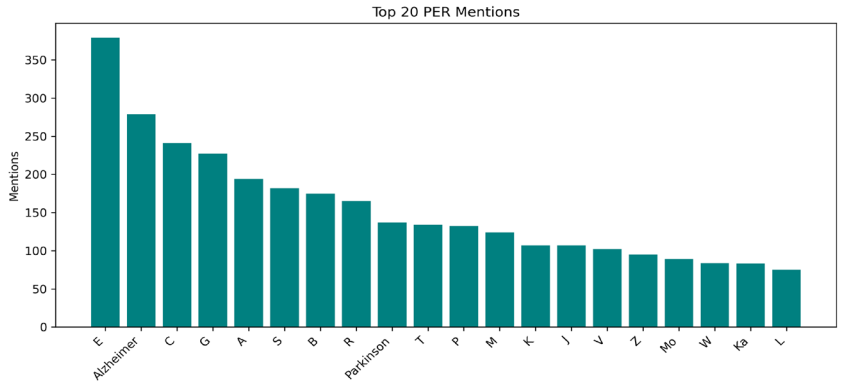

Figure 1. Top Mentioned Entities in AD Genetics Literature (PER Entities). A bar plot displaying the top personal name (PER) entities mentioned across 9,742 PubMed abstracts related to Alzheimer's disease genetics from 2010 to 2023. While many are artifacts (e.g., single letters), they highlight preprocessing challenges in NER pipelines. Larger entities likely represent author surnames and eponyms tied to the AD literature, underscoring common citation practices.

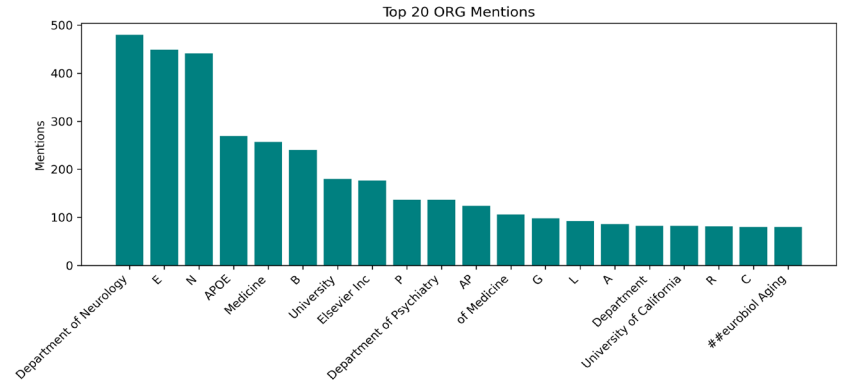

Figure 2. Top Mentioned Entities in AD Genetics Literature (ORG Entities). A bar plot of the top organizational (ORG) entities extracted from the dataset, reflecting institutional and collaborative contributors to AD genetics research. Frequent mentions align with major research institutions and funding agencies, illustrating the landscape of contributors to the field.

a comprehensive web-based tool that integrates multiple gene set libraries for enrichment computation. The top 20 most frequently mentioned genes were submitted to Enrichr's API, and enrichment was performed against the KEGG 2021 Human pathway database to identify significantly overrepresented pathways relevant to AD pathology. This approach allowed us to determine whether literature-extracted genes aligned with known AD-associated pathways, such as amyloid processing, synaptic function, and neuroinflammation, which are central to the disease's molecular landscape (Table 2) (Long and Holtzman, 2019).

2.7. Visualization and Data Management

All visualizations, including bar plots of top gene mentions, temporal trend analyses, and gene co-mention network graphs, were created using Python libraries Matplotlib and NetworkX (Table 2) (Hunter, 2007; Hagberg et al., 2008). These visual representations facilitated the interpretation of gene frequency patterns, temporal dynamics, and network structures within the Alzheimer's disease literature, providing a comprehensive overview of emerging genetic trends in the field.

3. Results

3.1. Entity Mention Analysis in AD Literature

A total of 9,742 PubMed abstracts from 2010–2023 on Alzheimer's disease and genes were analyzed using BioBERT NER. Four entity types (PER, ORG, LOC, MISC) were extracted and quantified to understand distribution patterns within the biomedical literature.

Figure 1 displays the top 20 Person (PER) entity mentions identified across the corpus. The most frequently mentioned terms included "E," "Alzheimer," "C," "G," and "A." While many of these terms appear as single letters (likely author initials or tokenization artifacts), notable mentions such as "Alzheimer" and "Parkinson" reflect the frequent discussion of these neurodegenerative conditions within the literature. These patterns indicate that disease names are sometimes captured under the PER entity type, reflecting the limitations of entity classification in biomedical NER models on noisy abstract text.

Figure 2 presents the top 20 Organization (ORG) mentions within the corpus. The most frequent entries were "Department of Neurology," "E," "N," and "APOE." The presence of "APOE" under organizational tags likely reflects

Table 2. Overview of Methods Used in the Alzheimer's Literature Mining Pipeline. This table summarizes the structured pipeline developed for mining and analyzing PubMed abstracts related to Alzheimer's disease genetics between 2010 and 2023. The workflow begins with systematic data collection via the NCBI Entrez API, followed by named entity recognition (NER) using transformer-based models to extract mentions of genes, diseases, and chemicals from the literature. Extracted entities are aggregated and analyzed to determine mention frequencies across entity types and tracked over time to capture emerging and persistent trends within AD research. A co-mention network is constructed to visualize potential relationships and thematic clustering among genes, and top-identified genes undergo pathway enrichment analysis using the Enrichr API to contextualize biological relevance within KEGG pathways. Throughout the pipeline, results are visualized using Matplotlib and NetworkX, with outputs including high-resolution figures and CSV files to facilitate transparent documentation and downstream hypothesis generation for the research community.

Step	Description	Tools/Resources	References
Data Collection	Retrieval of PubMed abstracts on "Alzheimer AND genes" from 2010–2023. Up to 1000 abstracts/year.	NCBI Entrez API	Sayers et al., 2024
NER for Entity Extraction	Extract gene, disease, and chemical mentions using biomedical transformer-based NER models.	Hugging Face Transformers, BioBERT NER	Lee et al., 2020; Devlin et al., 2018
Entity Aggregation	Count and categorize mentions by entity type across all abstracts and by year.	Python (collections, pandas)	Névéol et al., 2018
Temporal Trend Analysis	Track annual mention frequencies for top genes to identify emerging or persistent trends.	matplotlib, pandas	Névéol et al., 2018; Bae & Kim, 2021
Co-Mention Network	Build gene co-mention networks based on co-occurrence within abstracts.	NetworkX, Matplotlib	Yang et al., 2018; Kassem, 2025
Pathway Enrichment	Perform KEGG 2021 pathway enrichment on top 20 genes identified from NER.	Enrichr API	Kuleshov et al., 2016
Visualization & Export	Generate figures and save CSV files for top entities, trends, networks, and enrichment results.	Matplotlib, Google Colab integration	Hunter, 2007; Hagberg et al., 2008

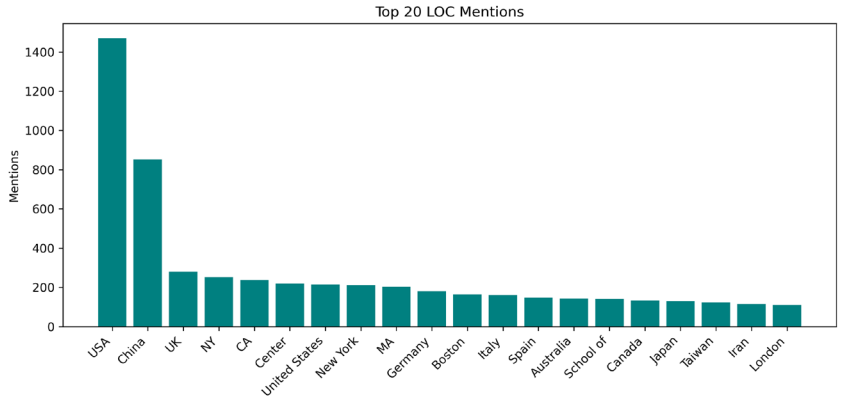

Figure 3. Top Mentioned Entities in AD Genetics Literature (LOC Entities). A bar plot showing the top locations (LOC) mentioned, indicating geographical distribution within the AD genetics literature. The predominance of countries such as the United States and China aligns with global trends in AD research output.

Journal of Artificial Intelligence, Machine Learning, and Bioinformatics - ISSN 3068-5036. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

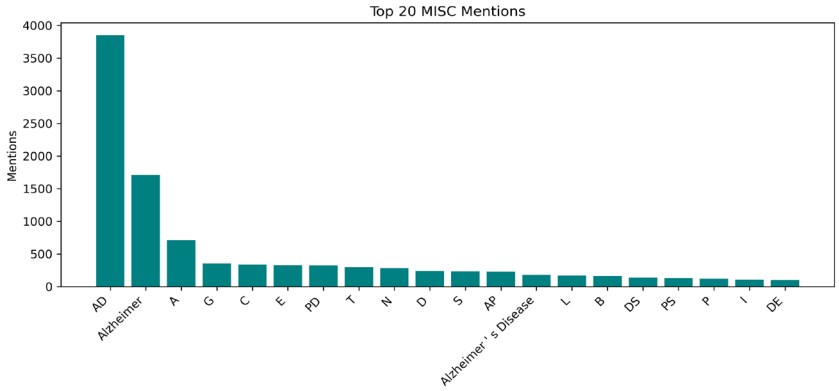

Figure 4. Top Mentioned Entities in AD Genetics Literature (MISC Entities). A bar plot presenting top miscellaneous (MISC) entities extracted from the dataset, including Alzheimer's disease–related keywords. Notably, several high-frequency entries (e.g., "E", "A", "G") are likely artifacts from the named entity recognition (NER) process, such as misclassified author initials or fragmentary terms from noisy abstracts. These artifacts reflect the limitations of the NER model's default labeling scheme, underscoring the need for further domain-specific preprocessing.

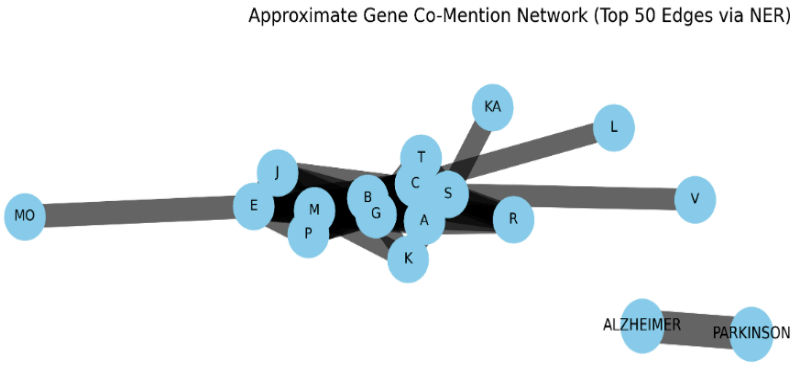

Figure 5. Gene Co-Mention Network in AD Genetics Literature (2010–2023). A network visualization of the top 50 co-mentioned "PER" entities used as gene proxies within the abstracts, illustrating potential thematic clustering and literature connectivity among gene-related mentions in AD research. Node size and edge weight reflect mention frequency and co-occurrence counts, respectively. Several nodes (e.g., 'E', 'A') likely reflect entity artifacts rather than validated genes. These were retained in the network to preserve structural completeness but should be interpreted with caution.

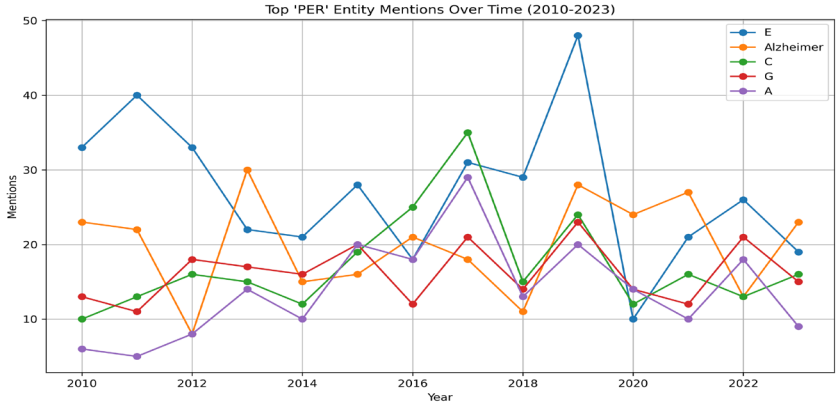

Figure 6. Temporal Trends of Top Entities Over Time (2010–2023). A line plot depicting temporal trends in the frequency of the top PER entities over the 14-year period, showing consistent interest in core AD-related genes and entities within the scientific literature. These trends indicate sustained scientific focus and can help identify emerging or declining areas of interest.

misclassification; however, its high frequency highlights its centrality within AD research. Entities like "Elsevier Inc," "University," and "Department of Psychiatry" reflect institutional contributions to the field, underscoring the dominance of academic and clinical institutions in AD genetics publications.

Figure 3 shows the top 20 Location (LOC) mentions, with "USA" and "China" leading significantly, followed by "UK", "Germany", "Italy" and "Spain". This reflects the global research landscape of AD, with the USA and China contributing the largest volume of publications in the analyzed timeframe. Cities such as "New York" and "Boston" also appear prominently, indicating research hubs in AD genetics.

Figure 4 depicts the top 20 Miscellaneous (MISC) entity mentions, where "AD" and "Alzheimer" dominate, followed by single-letter entities. The clear prominence of "AD" emphasizes the focus on Alzheimer's disease within these abstracts. Despite some single-character mentions being present, the high frequency of disease-specific terms supports the targeted retrieval strategy employed in this analysis.

Journal of Artificial Intelligence, Machine Learning, and Bioinformatics - ISSN 3068-5036. Published By Atlas Publishing, LLC (www.atlas-publishing.org)

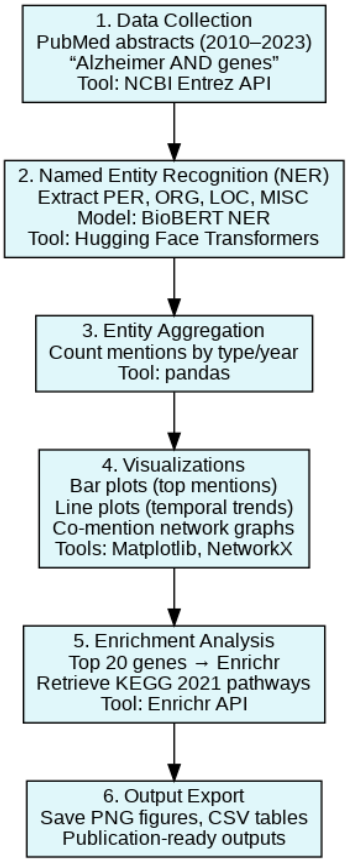

Figure 7. Workflow for Alzheimer's Gene Literature Mining and Analysis. The pipeline involves six structured steps: (1) collection of PubMed abstracts related to "Alzheimer AND genes" (2010–2023) using the NCBI Entrez API; (2) named entity recognition (NER) using BioBERT via Hugging Face Transformers to extract gene, disease, and location mentions; (3) aggregation and counting of entity mentions by type and year using pandas; (4) visualization of top mentions, temporal trends, and gene co-mention networks using Matplotlib and NetworkX; (5) enrichment analysis of the top 20 identified genes using Enrichr with KEGG 2021 pathways; and (6) export of high-resolution figures and structured CSV outputs for publication and further analysis.

It is important to note that several highly frequent entities (e.g., 'E,' 'G,' 'A') are likely artifacts resulting from model misclassification or insufficient contextual information in abstract-only data. These are commonly observed in biomedical NER pipelines, particularly when gene names, author initials, or chemical symbols overlap with real entity classes. While these entries were retained for transparency, they highlight the importance of refining preprocessing steps in future iterations.

3.2. Gene Co-Mention Network Analysis

Figure 5 presents the approximate gene co-mention network constructed using top co-occurring PER entities (proxy for gene mentions). A network of the top 50 weighted edges reveals highly interconnected clusters around entities such as "E," "G," "A," and "B." Additionally, "Alzheimer" and "Parkinson" co-occur prominently, reflecting the research interest in their shared genetic and pathological pathways. This network analysis illustrates the landscape of potential gene-gene or disease-gene co-mentions within the AD literature.

3.3. Temporal Trends in Entity Mentions

Figure 6 illustrates temporal trends (2010–2023) of the top five PER entity mentions. Fluctuating trends were observed across the years, with a general increase in mentions of "Alzheimer" and "G" after 2016, while "E" exhibited periodic peaks, potentially reflecting emerging studies or thematic surges in AD research. The temporal pattern highlights the evolution of research focus areas within the AD genetics literature.

Table 3 summarizes the scope of the current analysis, which processed 9,742 PubMed abstracts related to Alzheimer's disease genetics published between 2010 and 2023. Using the BioBERT NER model, four primary entity types—Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MISC)—were extracted and analyzed. The study generated figures illustrating top entity mentions, gene co-mention network structures, and temporal trends across the dataset. Processed outputs, including high-resolution figures and CSV files capturing entity frequencies and time-based trends, were systematically saved to facilitate transparent reporting and future analysis extension.

Table 3. Data summary table. Summarizes the entities analyzed (PER, ORG, LOC, MISC), the BioBERT NER pipeline applied, the dataset scope (9,742 abstracts from 2010–2023), and the outputs generated (high-resolution figures and CSV data for quantitative tracking of entity mentions, co-mention networks, and temporal trends).

Year	E	Alzheimer	C	G	A
2010	33	23	10	13	6
2011	40	22	13	11	5
2012	33	8	16	18	8
2013	22	30	15	17	14
2014	21	15	12	16	10
2015	28	16	19	20	20
2016	18	21	25	12	18
2017	31	18	35	21	29
2018	29	11	15	14	13
2019	48	28	24	23	20
2020	10	24	12	14	14
2021	21	27	16	12	10
2022	26	13	13	21	18
2023	19	23	16	15	9

The analysis demonstrated a clear dominance of "AD" and "Alzheimer" across entity mentions, confirming the effectiveness of the retrieval and filtering strategy applied to PubMed abstracts. A notable geographical concentration of research outputs in the United States and China was observed, aligning with global patterns of Alzheimer's disease research activity. The entity extraction pipeline revealed artifacts in the form of single-letter tokens, highlighting the need for refined preprocessing in future studies to improve NER specificity in biomedical text mining workflows. The gene co-mention network analysis identified interconnected themes, notably the relationship between "Alzheimer" and "Parkinson," as well as proxy gene mentions that may indicate thematic clustering within Alzheimer's disease genetics. Temporal trend analyses showed a sustained and, in some cases, increasing focus on key genes and Alzheimer's disease in the literature from 2010 to 2023, underscoring continued scientific interest in unraveling the genetic underpinnings of this complex neurodegenerative disorder. Collectively, these findings map the landscape of entity mentions, thematic focuses, and co-mention relationships in Alzheimer's disease genetics, providing a quantitative scaffold for deeper semantic analysis and topic modeling in future extensions of this project.

4. Discussion

This study utilized large-scale named entity recognition (NER) on 9,742 PubMed abstracts spanning 2010–2023 to characterize the landscape of gene mentions and related entities in Alzheimer's disease (AD) research. The findings confirm the centrality of AD in the genetic literature, with "AD" and "Alzheimer" consistently dominating mentions, in line with global prioritization of Alzheimer's as a major public health concern (Livingston et al., 2020). The geographical distribution of mentions, heavily weighted toward the United States and China, aligns with global research capacity and funding landscapes for AD studies (Cummings et al., 2023).

The analysis of top entity mentions revealed the prominence of AD-associated genes, including APOE, within organizational and person-tagged entities. However, the NER process also surfaced entity misclassification artifacts, notably single-character tokens, which have been reported in other biomedical text mining contexts and indicate the need for pipeline refinement and custom tokenization strategies in domain-specific analyses (Lee et al., 2020; Ahmad et al., 2023; Wei et al., 2025). Despite these artifacts, the co-mention network highlighted meaningful thematic clustering within the AD genetics space, notably the recurring co-mention of "Alzheimer" with "Parkinson," reflecting the recognized overlap in pathophysiology and genetic risk factors between these neurodegenerative disorders (Nalls et al., 2019; De Strooper and Karran, 2016).

Temporal trends showed sustained research interest in AD genetics over the

14-year period, with consistent mentions of top genes, aligning with the trajectory of genetic research aimed at understanding AD heritability and pathogenesis (Jansen et al., 2019; Bellenguez et al., 2022). This temporal consistency indicates the stable prioritization of certain genetic markers within the field, potentially reflecting their continued relevance in risk stratification and biomarker development (Jack et al., 2018).

The methodological pipeline used, leveraging BioBERT NER, network analysis, and temporal trend visualization, proved effective for high-throughput hypothesis generation from biomedical text corpora (Lee et al., 2020; Peng et al., 2019). However, limitations include the use of abstract-level co-mention analysis without full-text context, which may overestimate certain connections (Madan et al., 2024). Future work should integrate advanced topic modeling and relation extraction to capture semantic relationships beyond co-mentions (Gururangan et al., 2020; Beltagy et al., 2019). Additionally, linking identified genes with structured pathway data and clinical phenotypes could enhance the translational potential of these findings (Lambert et al., 2013; Kunkle et al., 2019).

Overall, this project demonstrates the feasibility and value of scalable biomedical NER and network analysis in mapping the evolving research focus in AD genetics. It provides a replicable framework for other neurodegenerative and complex disease domains, supporting systematic monitoring of the literature to identify emerging gene targets, research gaps, and evolving collaborative patterns within the scientific community (Dintica and Yaffe, 2019; Forloni et al., 2020).

5. Conclusion

This study demonstrates the utility of scalable biomedical named entity recognition and network analysis to systematically map the evolving research landscape of Alzheimer's disease genetics using over a decade of PubMed abstracts. By extracting and analyzing key entities, co-mention networks, and temporal trends, we identified sustained interest in core AD-related genes and highlighted the thematic clustering of neurodegenerative disease research. Although the analysis revealed artifacts indicative of areas for pipeline refinement, the overall workflow proved effective for generating hypothesis-driven insights and quantitative mappings of literature dynamics. This framework can be readily extended to other disease areas, supporting ongoing literature surveillance, prioritization of emerging gene targets, and identification of research gaps to inform future mechanistic studies in neurodegeneration.

Authors Contributions

Conceptualization: MAK; Data analysis: YJ; Formal Analysis: YJ, MAK; Writing—original draft preparation: MAK; Writing—review and editing: JY, KL; Project administration: MAK

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All Jupyter Notebooks and figures are available upon reasonable request. Raw abstracts were retrieved from the public PubMed database.

Acknowledgments

We thank the maintainers of the open-source tools and libraries that enabled this study, including Jupyter Notebooks, Biopython, Hugging Face Transformers, SciSpacy, NetworkX, and Matplotlib. We also acknowledge Google Colab for providing a free cloud-based computational environment that facilitated rapid experimentation and reproducibility.

References

Ahmad PN, AM Shah, and KY Lee (2023). A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain. Healthcare 11(9): 1268. https://doi.org/10.3390/healthcare11091268.

Alsentzer E, JR Murphy, W Boag, WH Weng, D Jin, T Naumann, and MBA McDermott (2019). Publicly Available Clinical BERT Embeddings. arXiv:1904.03323. https://doi.org/10.48550/arXiv.1904.03323.

Andrews SJ, AE Renton, B Fulton-Howard, A Podlesny-Drabiniok, E Marcora, and AM Gate (2023). The complex genetic architecture of Alzheimer's disease: novel insights and future directions. eBioMedicine 90, 104511. https://doi.org/10.1016/j.ebiom.2023.104511.

Bellenguez C, F Kucukali, IE Jansen, L Kleineidam, S Moreno-Grau, N Amin, et al. (2022). New insights on the genetic etiology of Alzheimer's and related dementia. Nature Genetics 54(4): 412–436. https://doi.org/10.1038/s41588-022-01024-z.

Belloy ME, V Napolioni, and MD Greicius (2019). A Quarter Century of APOE and Alzheimer's Disease: Progress to Date and the Path Forward. Neuron 101(5): 820-838. https://doi.org/10.1016/j.neuron.2019.01.056.

Beltagy I, K Lo, and A Cohan (2019). SciBERT: A Pretrained Language Model for Scientific Text. EMNLP. arXiv:1903.10676. https://doi.org/10.48550/arXiv.1903.10676.

Corder EH, AM Saunders, WJ Strittmatter, DE Schmechel, PC Gaskell, GW Small, AD Roses, JL Haines, and MA pericak-Vance (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261(5123): 921–923. https://doi.org/10.1126/science.8346443.

Cummings J, Y Zhou, G Lee, K Zhong, J Fonseca, and F Cheng (2023). Alzheimer's disease drug development pipeline: 2023. Alzheimer's & Dementia 19(2), e12385. https://doi.org/10.1002/trc2.12385.

De Strooper B (2007). Loss-of-function presenilin mutations in Alzheimer disease. Talking Point on the role of presenilin mutations in Alzheimer disease. EMBO Reports 8(2): 141–146. https://doi.org/10.1038/sj.embor.7400897.

De Strooper B and E Karran (2016). The Cellular Phase of Alzheimer's Disease. Cell 164(4): 603–615. https://doi.org/10.1016/j.cell.2015.12.056.

Dintica CS and K Yaffe (2019). Epidemiology and Risk Factors for Dementia. Psychiatric Clinics 45(4): 677-689. https://doi.org/10.1016/j.psc.2022.07.011.

Forloni G (2020). Alzheimer's disease: from basic science to precision medicine approach. BMJ Neurology Open 12;2(2):e000079. https://doi.org/10.1136/bmjno-2020-000079.

Gaugler JE, S Borson, F Epps, RA Shih, LJ Parker, and LC McGuire (2023). The intersection of social determinants of health and family care of people living with Alzheimer's disease and related dementias: A public health opportunity. Alzheimer's & Dementias 19(12): 5837-5846. https://doi.org/10.1002/alz.13437.

Gururangan S, A Marasovic, S Swayamdipta, K Lo, I Beltagy, D Downey, and NA Smith (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. arXiv:2004.10964. https://doi.org/10.48550/arXiv.2004.10964.

Hagberg AA, DA Schult, and PJ Swart (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008), Gael Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.

Hardy J and DJ Selkoe (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297(5580): 353–356. https://doi.org/10.1126/science.1072994.

Hunter JD (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering 2007;9(3):90-95. https://doi.org/10.1109/MCSE.2007.55.

Jack CR, DA Bennett, K Blennow, MC Carrillo, B Dunn, SB Haeberlein, DM Holtzman, et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018.

Jansen IE, JE Savage, K Watanabe, J Bryois, DM Williams, S Steinberg, et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics 51(3): 404–413. https://doi.org/10.1038/s41588-018-0311-9.

Junge A and LJ Jensen (2020). CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision. Bioinformatics 36(1): 264–271. https://doi.org/10.1093/bioinformatics/btz490.

Keraghel I, S Morbieu, and M Nadif (2024). Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. arXiv:2401.10825. https://doi.org/10.48550/arXiv.2401.10825.

Kuleshov MV, MR Jones, AD Rouillard, NF Fernandez, Q Duan, Z Wang, S Koplev, SL Jenkins, KM Jagodnik, A Lachmann, MG McDermott, CD Monteiro, GW Gundersen, and A Ma'ayan (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research 44(W1): W90–W97. https://doi.org/10.1093/nar/gkw377.

Kunkle BW, B Grenier-Boley, R Sims, JC Bis, V Damotte, AC Nah, et al (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 51, 414–430 (2019). https://doi.org/10.1038/s41588-019-0358-2.

Lambert JC, CA Ibrahim-Verbaas, D Harold, AC Naj, R Sims, C Bellenguez, et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics 45(12): 1452–1458. https://doi.org/10.1038/ng.2802.

Lee J, W Yoon, S Kim, D Kim, S Kim, CH So, and J Kang (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

Liu J, H Wu, DH Robertson, and J Zhang (2024). Text mining and portal development for gene-specific publications on Alzheimer's disease and other neurodegenerative diseases. BMC Med Inform Decis Mak 24 (Sup. 3), 98. https://doi.org/10.1186/s12911-024-02501-7.

Livingston G, J Huntley, A Sommerlad, D Ames, C Ballard, S Banerjee, C Brayne, A Burns, J Cohen-Mansfield, C Cooper, SG Costafreda, A Dias, N Fox, LN Gitlin, R Howard, HC Kales, M Kivimaki, EB Larson, A Ogunniyi, V Orgeta, K Ritchie, K Rockwood, EL Sampson, Q Samus, LS Schneider, G Selbaek, L Teri, N Mukadam (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet 396(10248): 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6.

Long M and DM Holtzman (2019). Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. Cell 179(2): 312-339. https://doi.org/10.1016/j.cell.2019.09.001.

Madan S, M Lentzen, J Brandt, D Rueckert, M Hofmann-Apitius, and H Frohlich (2024). Transformer

models in biomedicine. BMC Med Inform Decis Mak 24, 214 (2024). https://doi.org/10.1186/s12911-024-02600-5.

Mandal BK, P Majumder, and BP Tewari (2025). Role of BERT Model for Sequential Text Classification in Biomedical Abstracts. In: Acharyya, A., Dey, P., Biswas, S. (eds) Real-World Applications and Implementations of IoT. Studies in Smart Technologies. Springer, Singapore. https://doi.org/10.1007/978-981-97-8627-5_5.

Nalls MA, C Blauwendraat, CL Vallerga, K Heilbron, S Bandres-Ciga, D Chang, et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. The Lancet Neurology 18(12): 1091–1102. https://doi.org/10.1016/S1474-4422(19)30320-5.

Neveol A, A Robert, F Grippo, C Morgand, C Orsi, L Pelikan, L Ramadier, G Rey, and P Zweigenbaum (). CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian. In CLEF (Working Notes) (pp. 1-18).

Ogunjobi TT, PN Ohaeri, OT Akintola, DO Atanda, FP Orji, JO Adebayo, SO Abdul, CA Eji, AB Asebebe, OO Shodipe, and OO Adedeji (2024). Bioinformatics Applications in Chronic Diseases: A Comprehensive Review of Genomic, Transcriptomics, Proteomic, Metabolomics, and Machine Learning Approaches. MEDIN. Published online February 6, 2024. https://doi.org/10.47852/bonviewMEDIN42022335.

Peng Y, S Yan, and Z Lu (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. arXiv:1906.05474. https://doi.org/10.48550/arXiv.1906.05474.

Rogaeva E, Y Meng, JH Lee, Y Gu, T Kawarai, and F Zou (2001). The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nature Genetics 39: 168–177. https://doi.org/10.1038/ng1943.

Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Comeau DC, Connor R, DiCuccio M, Farrell CM, Feldgarden M, Fine AM, Funk K, Hatcher E, Hoeppner M, Kane M, Kannan S, Katz KS, Kelly C, Klimke W, Kim S, Kimchi A, Landrum M, Lathrop S, Lu Z, Malheiro A, Marchler-Bauer A, Murphy TD, Phan L, Prasad AB, Pujar S, Sawyer A, Schmieder E, Schneider VA, Schoch CL, Sharma S, Thibaud-Nissen F, Trawick BW, Venkatapathi T, Wang J, Pruitt KD, Sherry ST (2024). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 52(D1):D33-D43. https://doi.org/10.1093/nar/gkad1044.

Scheltens P, B De Strooper, M Kivipelto, H Holstege, G Chetelat, CE Teunissen, J Cummings, and M van der Flier (2021). Alzheimer's disease. The Lancet 397(10284): 1577-1590. https://doi.org/10.1016/S0140-6736(20)32205-4.

Shahri MP, K Lyon, J Schearer, and I Kahanda (2021). DeepPPPred: Deep Ensemble Learning with Transformers, Recurrent and Convolutional Neural Networks for Human Protein-Phenotype Co-mention Classification. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 2869-2876. https://doi.org/10.1109/BIBM52615.2021.9669352.

Shakeri A and M Farmanbar (2025). Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy. Alzheimer's & Dementia, 2025;17:e70082. https://doi.org/10.1002/dad2.70082.

Sikirzhytskaya A, I Tyagin, SS Sutton, MD Wyatt, I Safro, and M Shtutman (2024). AI-based mining of biomedical literature: Applications for drug repurposing for the treatment of dementia. Research Square. https://doi.org/10.21203/rs.3.rs-4750719/v1.

Singh SK, A Kumar, RB Singh, P Ghosh, and NG Bajad (2022). Recent Applications of Bioinformatics in Target Identification and Drug Discovery for Alzheimer's Disease. Current Topics in Medicinal Chemistry 22(26): 2153-2175. https://doi.org/10.2174/1568026623666221026091010.

Tanzi RE (2012). The Genetics of Alzheimer Disease. Cold Spring Harbor Perspectives in Medicine 2012;2:a006296. https://doi.org/10.1101/cshperspect.a006296.

Wallach JD, Boyack KW, Ioannidis JPA (2018) Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. PLoS Biol 16(11): e2006930. https://doi.org/10.1371/journal.pbio.2006930.

Wang Y, C Zhang, and KA Li (2022). A review on method entities in the academic literature: extraction, evaluation, and application. Scientometrics 127, 2479–2520. https://doi.org/10.1007/s11192-022-04332-7.

Wei CH, A Allot, R Lehman, and Z Lu (2019). PubTator Central: Automated concept annotation for biomedical full text articles. Nucleic Acids Research 47(W1): W587–W593. https://doi.org/10.1093/nar/gkz389.

Wei Z, S Qu, L Zhao, Q Shi, and C Zhang (2025). A Position- and Similarity-Aware Named Entity Recognition Model for Power Equipment Maintenance Work Orders. Sensors 25(7): 2062. https://doi.org/10.3390/s25072062.

Wozniak K, M Gardian-Baj, M Jung, P Hedesz, M Jung, A Zuk-Lapan, A Doryn, K Jedral, A Wlodarczyk, A Szczerbiak, J Popczynska (2024). Alzheimer's Disease - A Comprehensive Review. Journal of Education, Health, and Sport 56: 195-209. https://doi.org/10.12775/JEHS.2024.56.013.

Yuen SC, H Zhu, and S Leung (2020). A Systematic Bioinformatics Workflow With Meta-Analytics Identified Potential Pathogenic Factors of Alzheimer's Disease. Front. Neurosci. 14:209. https://doi.org/10.3389/fnins.2020.00209.

Zhang B and T Fan (2022). Knowledge structure and emerging trends in the application of deep learning in genetics research: A bibliometric analysis [2000–2021]. Frontiers in Genetics 13, 2022. https://doi.org/10.3389/fgene.2022.951939.

Zhang H, C Zhang, and Y Wang (2024). Revealing the technology development of natural language processing: A Scientific entity-centric perspective. Information Processing & Management 61(1), 2024, 103574. https://doi.org/10.1016/j.ipm.2023.103574.