BiblioEngine: An AI-empowered Platform for Disease Genetic Knowledge Mining

Mengjia Wu[1 [0000−0003−3956−7808]], Yi Zhang[1 [0000−0002−7731−0301]], Hua Lin[2], Mark Grosser[2], Guangquan Zhang[1 [0000−0003−3960−0583]], and Jie Lu[1 [0000−0003−0690−4732]]

[1] University of Technology Sydney, Ultimo, New South Wales, Australia
mengjia.wu@uts.edu.au, yi.zhang@uts.edu.au, guangquan.zhang@uts.edu.au, jie.lu@uts.edu.au
[2] 23Strands, Pyrmont, New South Wales, Australia
mark.grosser@23strands.com, hua.lin@23strands.com

Abstract. Recent decades have seen significant advancements in contemporary genetic research with the aid of artificial intelligence (AI) techniques. However, researchers lack a comprehensive platform for fully exploiting these AI tools and conducting customized analyses. This paper introduces BiblioEngine, a literature analysis platform that helps researchers profile the research landscape and gain genetic insights into diseases. BiblioEngine integrates multiple AI-empowered data sources and employs heterogeneous network analysis to identify and emphasize genes and other biomedical entities for further investigation. Its effectiveness is demonstrated through a case study on stroke-related genetic research. Analysis with BiblioEngine uncovers valuable research intelligence and genetic insights. It provides a profile of leading research institutions and the knowledge landscape in the field. The gene co-occurrence map reveals frequent research of NOTCH3, prothrombotic factors, inflammatory cytokines, and other potential risk factors. The heterogeneous biomedical entity network analysis highlights infrequently studied genes and biomedical entities with potential significance for future stroke studies. In conclusion, BiblioEngine is a valuable tool enabling efficient navigation and comprehension of expanding biomedical knowledge from scientific literature, empowering researchers in their pursuit of disease-specific genetic knowledge.

Keywords: Artificial intelligence · Network analytics · Disease genetics.

1 Introduction

In recent decades, contemporary genetic research has witnessed notable advancements by leveraging Artificial Intelligence (AI) techniques. These techniques have enabled the analysis of large-scale, high-dimensional, and multi-modal data [8, 9], which particularly help identify the associations between diseases and gene mutations in the disease genetic context. Nevertheless, an important challenge persists in the absence of a comprehensive platform that aids biomedical researchers in

efficiently comprehending the explosive knowledge embedded within the rapidly accumulating scientific literature. To address this gap, this paper introduces BiblioEngine, a literature knowledge-mining platform that incorporates analytical techniques empowered by natural language processing (NLP) and network analytics. By harnessing multiple AI-empowered data sources, BiblioEngine equips the original research papers with research concepts and biomedical entities. Additionally, it employs a heterogeneous network analysis framework to unveil the significance and specificity of these biomedical entities, ultimately generating comprehensive rankings across four distinct categories.

This paper additionally presents a case study that investigates stroke genetic knowledge through literature analysis. First, it elucidates the research intelligence and knowledge landscape within the field, highlighting prominent institutions such as Inserm, Harvard University, and Karolinska Institutet, as well as the emerging influence of the University of Cambridge and Massachusetts General Hospital. Secondly, it unravels the genetic knowledge pertaining to stroke, identifying NOTCH3 as the most frequently studied gene. Additionally, it demonstrates that other genes tend to be studied in conjunction, forming three distinct groups: prothrombotic factors, inflammatory cytokines, and other potential risk factors. Through gene importance-specificity analysis, six genes are identified as meriting further investigation due to their current potential, albeit limited evidence. The overall rankings derived from the study highlight several biomedical entities that warrant further exploration regarding their associations with stroke.

Our major contributions are two-fold: Frist, it introduces a workflow that harmoniously integrates AI-empowered tools and network analytics to facilitate the efficient assimilation of disease-related genetic knowledge derived from scientific literature. Second, It presents an in-depth and comprehensive case study centered on stroke, elucidating the current research landscape in this area. Additionally, the case study highlights frequently studied genes and identifies potential genes and mutations that warrant further investigation.

The remainder of this paper is organized as follows: Section 2 delineates the methodology and workflow of BiblioEngine, providing a detailed explanation of its functioning. Section 3 presents the case study conducted on stroke, including the research findings and corresponding results. Finally, Section 4 concludes the paper by summarizing the key insights and discussing the limitations encountered during the study.

2 Platform Framework and Methodology

The framework and workflow of BiblioEngine are depicted in Figure 1. The platform operates by receiving research paper inputs from the PubMed database. These papers are then mapped to OpenAlex[3] and Pubtator[4] to retrieve standardized hierarchical research concepts and entities that are studied within the

[3] https://openalex.org/
[4] https://www.ncbi.nlm.nih.gov/research/pubtator/

research papers. Subsequently, the platform encompasses two distinct task trajectories. The first trajectory is dedicated to the discovery of research intelligence, and the second trajectory focuses on revealing gene knowledge derived from the literature via heterogeneous biomedical entity network analysis.

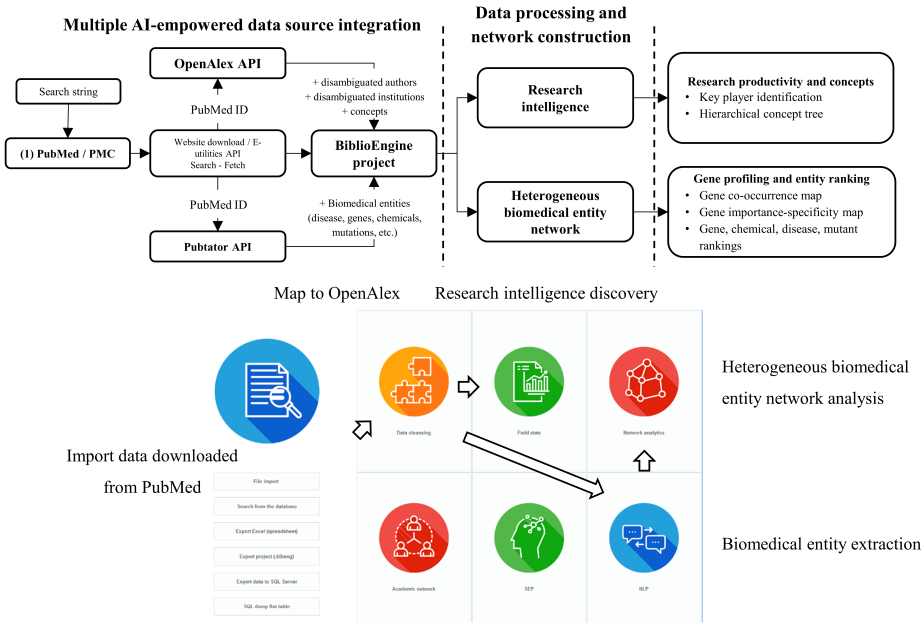

Fig. 1. Genetic knowledge mining framework

2.1 Multiple AI-empowered Data Source Integration

The BiblioEngine platform integrates two AI-empowered data sources, namely Pubtator and OpenAlex, to maximize the utilization of curated literature metadata provided by these repositories. OpenAlex offers hierarchical research concepts that reflect the underlying research content. Those concepts are derived from Wikipedia entries and are associated with research papers through a topic modeling process. Furthermore, the BiblioEngine platform integrates the Pubtator entity extraction function, which enables the identification of diverse biomedical entities from PubMed articles. These entities include diseases, genes, chemicals, mutations, cell lines, and more. The extracted entities are subsequently mapped to biomedical thesauri, enriching their semantic context. Moreover, the platform leverages corresponding entity databases to retrieve comprehensive information regarding the identified entities, including official symbols, names, mutation sites, and other relevant details.

2.2 Data Processing and Network Construction

The Pubtator API enables the extraction of biomedical entities, encompassing four primary categories: diseases, chemicals, genes, and mutations. In order to facilitate subsequent analysis, a weighted heterogeneous network is constructed, incorporating these four biomedical entity categories. The constructed network can be represented by $G = (V^K, E)$, K is the number of entity categories, V_i^m represents the mth node in the ith category, and the edges are attributed with the sentence co-occurrence frequency of two connected nodes.

2.3 Research Intelligence Discovery

Research intelligence involves capturing a comprehensive overview of the research landscape within a specific domain and identifying key entities associated with that domain. These entities may include individuals, institutions, and regions/countries. The identification of key players can be determined by examining the productivity of individuals or entities over time, considering factors such as the number of papers they have published. In terms of profiling the research landscape, the BiblioEngine platform leverages the research concepts obtained from OpenAlex. These concepts are utilized to construct a hierarchical topic structure that reflects the content and themes of papers published within the research domain. By analyzing this hierarchical structure, insights can be gained regarding the composition and organization of research within the field.

2.4 Entity Ranking and Gene Profiling

In our previous work [13], we introduced an entity ranking and gene profiling approach that focuses on identifying key biomedical entities and uncovering gene characteristics in relation to a specific target disease. This approach utilizes centrality measures and network-based intersection ratios as proxies to assess the significance and specificity of biomedical entities within the context of the disease. Specifically, degree, closeness, and betweenness centrality measures are employed to the nodes' ability to aggregate, disseminate, and transfer information within a network. The calculations are provided below:

$$Degree(V_i^m) = \frac{\sum_{j=1}^{K} \sum_{n=1}^{|V_j|} A_{V_i^m V_j^n}}{|V_K| - 1} \tag{1}$$

$$Closeness(V_i^m) = \frac{|V_K| - 1}{\sum_{j=1}^{K} \sum_{n=1}^{|V_j|} d_{V_i^m V_j^n}} \tag{2}$$

$$Betweenness(V_i^m) = \frac{2 \sum_{x,y=1}^{K} \sum_{a=1}^{|V_x|} \sum_{b=1}^{|V_y|} \frac{\sigma(V_x^a V_y^b)_{V_i^m}}{\sigma(V_x^a V_y^b)}}{(|V_K| - 1)(|V_K| - 2)} \tag{3}$$

where $|V_K|$ denotes the number of all K categories of nodes in the network and $|V_j|$ is the number of nodes in the jth category, $d_{V_i^m V_j^n}$ is the distance from

node V_i^m to node V_j^n, $\sigma(V_x^a V_y^b)$ is all shortest path counts from node V_x^a to V_y^b and $\sigma(V_x^a V_y^b)_{V_i^m}$ is the path counts that pass through node V_i^m.

In addition to the three centrality measures mentioned earlier, we also calculate an additional indicator known as the intersection ratio. This indicator is used to assess the specificity of biomedical entities in relation to the target disease. which is defined as follows:

$$Intersection\ ratio(V_i^m) = \frac{w(V_i^m, V_{disease}^t)}{\sum_{a=1}^{|V_disease|} w(V_i^m, V_{disease}^a)} \tag{4}$$

where $V_{disease}^t$ represents the node(s) of the target disease, and $w(V_i^m, V_{disease}^t)$ refers to the weight of the edge connecting V_i^m and $V_{disease}^t$.

To obtain comprehensive rankings for each category of biomedical entities, we utilize a non-dominated sorting algorithm [13] that combines the three centrality measures as proxies for entity importance. This algorithm also incorporates the intersection ratio later to generate more comprehensive rankings. The non-dominated sorting algorithm ensures that entities are ranked in a non-dominated order, considering their importance as well as their specificity.

3 Case Study: Stroke Genetic Research Analysis

To showcase the effectiveness of the BiblioEngine platform, we conducted a case study centered around stroke genetic investigation. Stroke is a prevalent cardiovascular disease with detrimental implications for individuals and public health but poses a complex genetic landscape that remains incompletely understood. Using the BiblioEngine platform, we generated results that shed light on the research landscape and conveyed genetic knowledge pertinent to stroke. To enhance the visual representation of these findings, we employed Xmind[5] software and the Circos visualization web application[6].

3.1 Data collection and processing

Using the search strategy described below in PubMed, 4,557 original research papers were retrieved from PubMed for further analysis.

((("stroke/genetics"[MeSH Terms] OR ("Stoke"[All Fields] AND "genom*"[All Fields])) AND "humans"[MeSH Terms]) NOT ("Review"[Publication Type] OR "systematic review"[Publication Type])) AND (humans[Filter])
Search date: 31/05/2023

In accordance with the workflow of the BiblioEngine platform, the research papers obtained were mapped to the OpenAlex database, and subsequently, biomedical entities were extracted using Pubtator. The fundamental statistics

[5] https://xmind.app/
[6] http://mkweb.bcgsc.ca/tableviewer/visualize/

regarding the 4,557 research papers and the concepts/entities mined from them are presented in Table 1. Continuing with the network construction process described in Section 2.2, a co-occurrence network of heterogeneous biomedical entities was established. This network comprises 6,150 nodes and 42,021 edges.

Table 1. Basic statistics of the 4,557 research papers

Data source	Field	Count	Top five instances
OpenAlex	Author	13,284	Martin Dichgans, Hugh S. Markus, James F. Meschia, Bradford B. Worrall, Rainer Malik
OpenAlex	Institution	2,980	Harvard University, University of Cambridge, Inserm, Ludwig-Maximilians-Universität München, Massachusetts General Hospital
OpenAlex	Concept	110	Genotype, allele, single-nucleotide polymorphism, ischemic stroke, leukoencephalopathy
Pubtator	Disease	950	Stroke, Ischemic stroke, hypertension, atherosclerosis, leukoencephalopathy
Pubtator	Gene	1,868	NOTCH3, MTHFR, APOE, ACE, IL6
Pubtator	Chemical	495	Cholesterol, lipid, homocysteine, triglyceride, cysteine
Pubtator	Mutant	2,405	rs1801133, rs1799963, rs1799983, rs1801131, rs1800795

3.2 Key players and knowledge landscape in stroke genetic studies

This section delivers the first trajectory of results. In Figure 2, the ranking changes of the top ten leading research institutions are depicted. The time period is divided into six windows, with the first window ending in 2000, followed by five 5-year gaps.

Notably, the University of Cambridge and Massachusetts General Hospital have experienced notable rises in their rankings over the past decade. Analysis of research papers originating from these institutions indicates that their improved rankings can be attributed to the high productivity of research teams led by Hugh S. Markus and Jonathan Rosand, respectively. Harvard University, Karolinska Institutet, and Ludwig-Maximilians-Universität München consistently maintain high rankings throughout the entire time span, demonstrating their enduring research leadership and diverse contributions within the field. The rankings of other institutions exhibit fluctuating trends, which may be influenced by factors such as the migration of key research teams or time gaps between research progress and publication.

Subsequently, we proceeded to extract the hierarchical relationships among all concepts and present the visualization of the hierarchical concept tree in Figure 3. This hierarchical concept tree provides a visual representation of the

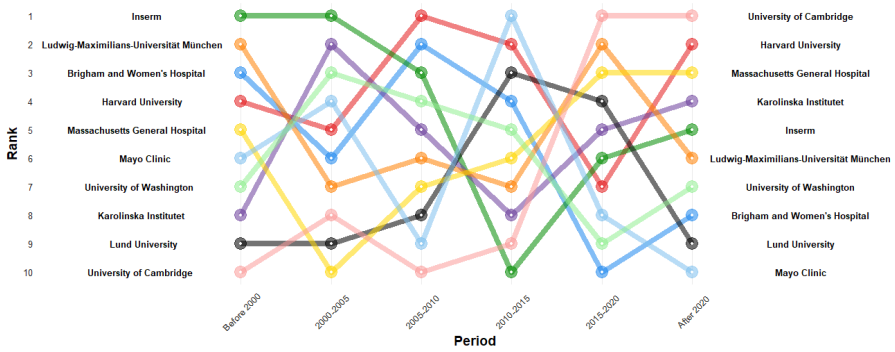

Fig. 2. Research institution ranking change in consecutive time windows

knowledge landscape pertaining to genetic studies of stroke. The numbers accompanying each concept indicate the count of research papers associated with that concept. To enhance clarity, we pruned the tree by removing branches that did not contain any concepts related to more than 100 research papers. By examining the hierarchical concept tree, we can observe that genetic studies of stroke encompass a wide range of research topics from the perspectives of basic medical sciences, clinical medicine, and public health. These topics encompass various indicators, conditions, research methods, and subjects that contribute to understanding the onset, progression, and prognosis of stroke. Furthermore, the branch related to biology reveals the emergence of computational biology and bioinformatics, signifying the increasing significance of data-driven genetic knowledge discovery in recent years. This development underscores the value and impact of integrating computational approaches in advancing genetic research on stroke.

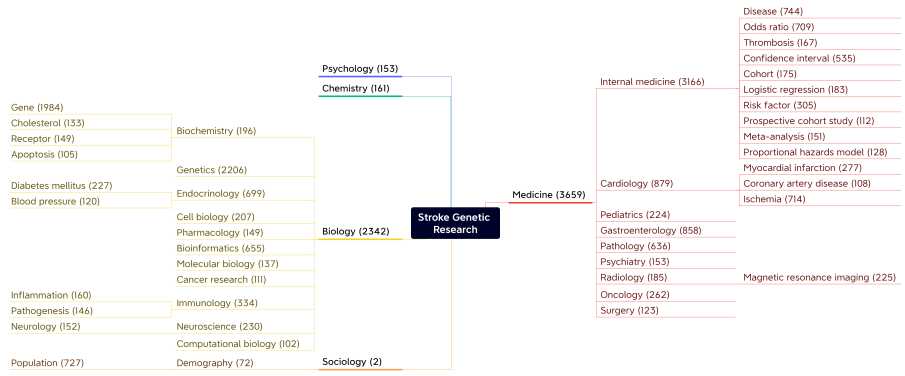

Fig. 3. Hierarchical concept tree of stroke genetic research

3.3 Gene maps and biomedical entity rankings

Gene co-occurrence map In order to provide a comprehensive overview of gene interactions from a literature-based standpoint, we present a gene co-occurrence map in Figure 4. This map displays the top 20 genes along with their co-occurrence links. The width of the ribbons connecting two genes represents the strength of their co-occurrence.

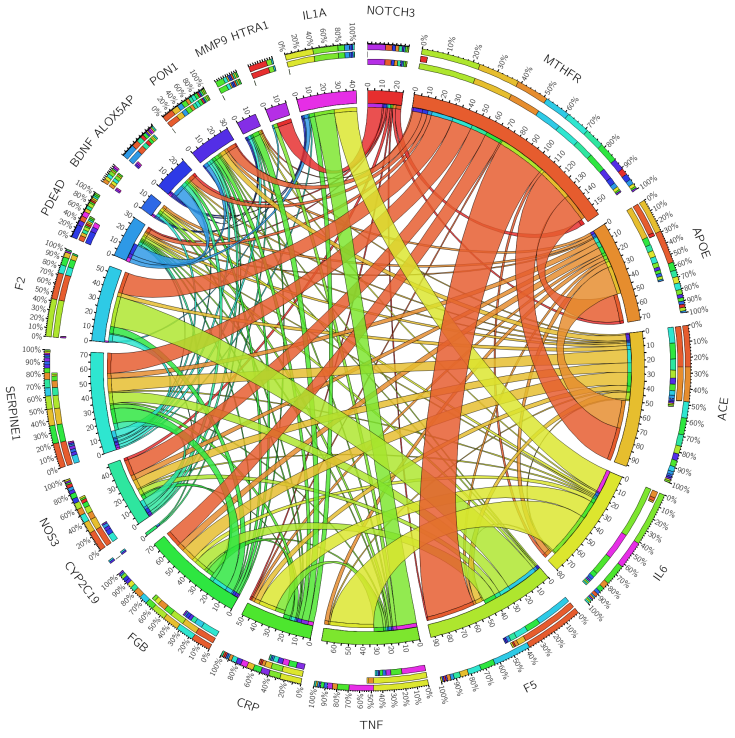

Fig. 4. Co-occurrence map of top 20 genes

By examining this gene co-occurrence map, we can gain insights into the potential interactions and relationships between different genes within the context of the studied literature. The visualization highlights interesting patterns regarding gene interactions in the studied literature. Specifically, the genes **MTHFR**, **APOE**, and **ACE** frequently appear together with other genes, suggesting potential associations or shared research interests. On the other hand, the gene **NOTCH3** stands out as being relatively studied alone, possibly due to its distinctive associations with CADASIL, a condition linked to stroke.

Further analysis of relevant papers provides insights into specific gene pairs: (1) Prothrombotic factors: The genes **MTHFR** and **F5** are commonly studied

together in genetic association studies due to their shared genetic variants, such as Factor V Leiden and C677T, which are known prothrombotic factors associated with increased stroke risks [2]. Additionally, mutations in the gene **F2** (prothrombin G20210) are frequently observed in such studies. (2) Inflammatory cytokine: The genes **IL6** and **TNF** play significant roles in inflammatory responses that occur in the brain, which can be triggered by tissue damage or injuries following a stroke [1]. Upregulation of these genes and the subsequent release of corresponding proteins contribute to the immune response and inflammatory cascade observed in the brain after a stroke. (3) Other possible risk factors: Mutations in the genes **ACE** and **APOE** are more commonly associated with heart diseases and Alzheimer's disease. However, their associations with stroke have been investigated, yielding inconsistent evidence [4]. These findings highlight the relationships and potential roles of these gene pairs in stroke and related conditions, providing valuable insights into ongoing research in the field.

Gene importance-specificity map To overcome the potential bias towards frequently studied genes and to identify genes that play significant but less explored roles in the onset and progression of stroke, we employed two indicators, importance and intersection ratio as introduced in Section 2.4, to profile all the identified genes. Figure 5 presents the gene profiling results, where the x-axis represents the centrality dominating score of genes, and the y-axis indicates the intersection ratio of genes specific to stroke.

By examining the gene profiling plot, we can identify genes that have lower frequencies but are more specifically associated with stroke. In addition to well-studied genes such as IL6, TNF, and NOS3, the map highlights a few genes that are relatively low-ranked but possess high dominating scores and specificity, denoted by red markers in Figure 5. To gain further insights into their associations with stroke, we delved into relevant research papers and summarized the findings in Table 3. The results reveal that mutations in these genes potentially have associations with different subtypes of stroke in diverse populations. However, it is important to note that the exploration of such associations is limited, necessitating additional case studies and experiments to establish and comprehend the underlying mechanisms. These genes warrant further investigation and may provide novel insights into the mechanisms and treatment of stroke.

Overall rankings of all biomedical entities Additionally, the dominating scores of the three centrality measures and the intersection ratio were computed for all entities, resulting in overall rankings across four categories. These rankings, presented in Table 3, provide a comprehensive integration of the four measurement dimensions. Notably, they highlight several emerging low-frequency entities across the four categories.

The overall rankings provide valuable insights into entities that warrant further investigation in future stroke studies. In the list of diseases, the majority are conditions, complications, or risk factors associated with stroke. The gene ranking reveals that the top two entities are both MicroRNAs, and the analysis

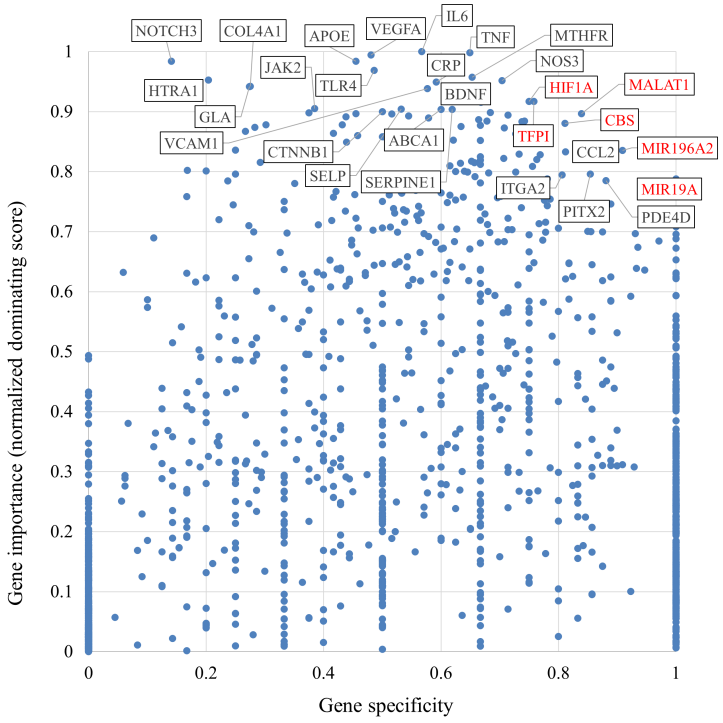

Fig. 5. Gene importance-specificity map

Table 2. Low-frequency genes highlighted by the gene importance-specificity map

Gene	Findings and evidence
HIF1A	A gender-based gene regulation comparison on stroke patients indicates that HIF1A is associated with female-specific stroke genes [11].
CBS	CBS is found to influence poststroke homocysteine metabolism which is associated with recurrent stroke [5].
TFPI	It was found that plasma TFPI level shows a significant effect as an environmental factor in addition to heritable factors [10].
MALAT1	Serum MALAT1 level and the mutant rs3200401 in MALAT1 are identified to be independent predictors of cerebral ischemic stroke [3].
MIR196A2	The combination of miR-146aG/-149T/-196a2C/-499G allele is found to be associated with the pathogenesis of ischemic stroke [6].
MIR19A	The expression of MIR19A is found to be significantly decreased in acute ischemic stroke patients [7].

Table 3. Overall rankings of four categories of biomedical entities

Disease	Gene	Chemical	Mutation
intracerebral hemorrhage	MIR19A	aspirin	rs2292832
coronary artery disease	MIR125A	homocysteine	rs6265
neurological disease	MAF	clopidogrel	rs710446
atherosclerosis	SHMT1	triglyceride	rs579459
venous thromboembolism	MIRLET7I	lipid	rs6843082
hypertension	ANXA3	polyacrylamide	rs243865
myocardial infarction	SLC22A4	technetium	rs1126579
atrial fibrillation	CST12P	ginkgolides	rs3200401
inflammation	IFNG	arachidonic acid	rs320
bleeding	FOS	methionine	rs28933697

of relevant papers suggests their potential as biomarkers for ischemic stroke [7]. Among the chemicals, ginkgolides emerge as a potential therapy for preventing stroke progression, although the evidence remains limited and conflicting [12]. Additionally, there is evidence suggesting that the level of methionine can possibly be associated with increased stroke risk. The mutation list highlights less frequently studied mutations, such as rs2292832 and rs579459, which may hold potential associations with stroke and thus require further investigation.

4 Conclusions and Further Study

This paper presents BiblioEngine, an biomedical literature analysis platform designed to aid researchers in efficiently and effectively extracting disease genetic knowledge from research papers. The platform integrates two AI-empowered data sources and employs a heterogeneous network analysis framework to facilitate biomedical entity ranking and gene prioritization. Its effectiveness is demonstrated by a case study on stroke genetic research analysis.

This platform is subject to certain limitations. Efforts are underway to enhance its functionalities in three key directions: First, the current network does not distinguish between positive and negative associations in gene and disease pairs. Future iterations of network construction will incorporate sentiment scores as edge attributes, enabling the indication of positive and negative associations. Second, The prediction of gene-disease associations holds significant value and will be prioritized as the next step in network analysis, treating it as a link prediction problem. Last, The interpretation of the obtained results can be further enhanced through real-world experimental validation, which may require ethical considerations [14]. To pursue this objective, collaboration with researchers specializing in this field is being actively sought.

Acknowledgements This work is supported by the Australian Research Council Linkage Project LP210100414.

References

1. Banerjee, I., Gupta, V., Ahmed, T., Faizaan, M., Agarwal, P., Ganesh, S.: Inflammatory system gene polymorphism and the risk of stroke: a case–control study in an indian population. Brain Research Bulletin **75**(1), 158–165 (2008)
2. Curry, C.J., Bhullar, S., Holmes, J., Delozier, C.D., Roeder, E.R., Hutchison, H.T.: Risk factors for perinatal arterial stroke: a study of 60 mother-child pairs. Pediatric Neurology **37**(2), 99–107 (2007)
3. Fathy, N., Kortam, M.A., Shaker, O.G., Sayed, N.H.: Long noncoding rnas malat1 and anril gene variants and the risk of cerebral ischemic stroke: an association study. ACS Chemical Neuroscience **12**(8), 1351–1362 (2021)
4. Gao, X., Yang, H., ZhiPing, T.: Association studies of genetic polymorphism, environmental factors and their interaction in ischemic stroke. Neuroscience Letters **398**(3), 172–177 (2006)
5. Hsu, F.C., Sides, E., Mychaleckyj, J., Worrall, B., Elias, G., Liu, Y., Chen, W.M., Coull, B., Toole, J., Rich, S., et al.: Transcobalamin 2 variant associated with poststroke homocysteine modifies recurrent stroke risk. Neurology **77**(16), 1543–1550 (2011)
6. Jeon, Y.J., Kim, O.J., Kim, S.Y., Oh, S.H., Oh, D., Kim, O.J., Shin, B.S., Kim, N.K.: Association of the mir-146a, mir-149, mir-196a2, and mir-499 polymorphisms with ischemic stroke and silent brain infarction risk. Arteriosclerosis, thrombosis, and vascular biology **33**(2), 420–430 (2013)
7. Jickling, G.C., Ander, B.P., Zhan, X., Noblett, D., Stamova, B., Liu, D.: microrna expression in peripheral blood cells following acute ischemic stroke and their predicted gene targets. PloS One **9**(6), e99283 (2014)
8. Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G.: Transfer learning using computational intelligence: A survey. Knowledge-Based Systems **80**, 14–23 (2015)
9. Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G.: Recommender system application developments: a survey. Decision Support Systems **74**, 12–32 (2015)
10. Nowak-Göttl, U., Langer, C., Bergs, S., Thedieck, S., Sträter, R., Stoll, M.: Genetics of hemostasis: differential effects of heritability and household components influencing lipid concentrations and clotting factor levels in 282 pediatric stroke families. Environmental Health Perspectives **116**(6), 839–843 (2008)
11. Tian, Y., Stamova, B., Jickling, G.C., Liu, D., Ander, B.P., Bushnell, C., Zhan, X., Davis, R.R., Verro, P., Pevec, W.C., et al.: Effects of gender on gene expression in the blood of ischemic stroke patients. Journal of Cerebral Blood Flow & Metabolism **32**(5), 780–791 (2012)
12. Wang, T.J., Wu, Z.Y., Yang, C.H., Cao, L., Wang, Z.Z., Cao, Z.Y., Yu, M.Y., Zhao, M.R., Zhang, C.F., Liu, W.J., et al.: Multiple mechanistic models reveal the neuroprotective effects of diterpene ginkgolides against astrocyte-mediated demyelination via the paf-pafr pathway. The American Journal of Chinese Medicine **50**(06), 1565–1597 (2022)
13. Wu, M., Zhang, Y., Zhang, G., Lu, J.: Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. Technological Forecasting and Social Change **164**, 120513 (2021)
14. Zhang, Y., Wu, M., Tian, G.Y., Zhang, G., Lu, J.: Ethics and privacy of artificial intelligence: Understandings from bibliometrics. Knowledge-Based Systems **222**, 106994 (2021)