Configurable Platform for Biomedical Literature Mining via Multimodal-Driven Extraction

Xinpan Yuan[1], Bozhao Li[1,2], Guihu Zhao[2,3](✉), Yueming Wang[4], Liujie Hua[1], Junhua Kuang[1], Jianguo Chen[5], Shaomin Xie[1], and Gan Li[1]

[1] School of Computer Science and Artificial Intelligence, Hunan University of Technology, Zhuzhou, Hunan 412007, China

[2] National Clinical Research Center for Geriatric Disorders, Department of Geriatrics, Xiangya Hospital & Center for Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, 410008, China
ghzhao@csu.edu.cn

[3] Bioinformatics Center, Furong Laboratory and Xiangya Hospital, Central South University, Changsha, Hunan, China

[4] Changsha Yahuilong Biological Technology Co., Ltd. Changsha, Hunan, China

[5] College of Computer and Data Science, Fuzhou University, Fuzhou, FuJian, China

Abstract. Biomedical literature serves as a critical repository for cutting-edge research achievements, encompassing substantial statistically validated biological knowledge. However, the dispersed storage and unstructured characteristics of such literature significantly hinder manual acquisition efficiency while increasing error susceptibility. To address these challenges, this study proposes an intelligent literature knowledge mining platform. Three core innovations distinguish this research: (1) The development of an extensible literature collection-parsing-structuring framework based on a "literature tree" architecture (*ECPS-LitTree*), which facilitates HTML dynamic report generation and full-cycle data management, offering a novel solution for cross-source heterogeneous literature knowledge aggregation; (2) The design of a configurable requirement customization framework (*CRC*) that combines named entity recognition (NER) technology with user-configurable mining templates to enable personalized knowledge extraction; (3) The implementation of an integrated online platform, providing comprehensive services including visual analytics, interactive search, and batch data export functionalities. Experimental validation demonstrates that the platform surpasses existing mainstream tools in literature retrieval success rate, processing efficiency, and knowledge extraction volume. The platform's flexible configurability exhibits broad applicability across multiple biomedical domains, offering researchers a reliable intelligent tool for knowledge discovery. The Configurable Platform is publicly and freely accessible at https://medseeker.genemed.tech/.

Keywords: Biomedical Literature · Document Parsing · Requirement Configuration · Named Entity Recognition · Interactive Website.

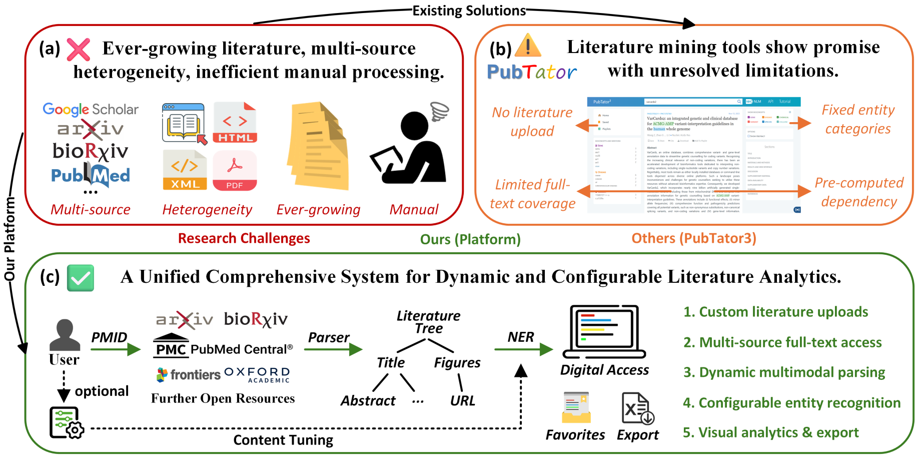


Fig. 1. Research Motivation: (a) Literature Mining Challenges; (b) Limitations of Existing Solutions (e.g., PubTator3); (c) Proposed Technical Framework with Key Technical Features (1-5).

1 Introduction

Biomedical literature forms the core knowledge base for biological and clinical sciences, providing validated cross-domain evidence [1]. The exponential publication growth—particularly in genomics [2]—presents two major challenges: inefficient integration of massive data and complexity in extracting knowledge from heterogeneous sources (see Fig. 1(a)). Precision medicine tasks like clinical variant pathogenicity classification [3] require automated tools to reliably extract evidence from dispersed literature. Concurrently, emerging high-impact discoveries demand intelligent systems capable of processing multimodal, cross-domain biomedical knowledge.

In genetic variant assessment, standards such as those established by the American College of Medical Genetics and Genomics (ACMG) provide critical frameworks for evidence-based clinical genetics practices [5,6]. However, applying criteria like PS2/PS3 requires laborious cross-referencing of multiple databases [7,15,16,17] and exhaustive analysis of peer-reviewed English literature. This manual workflow introduces subjective biases, frequently leading to inconsistent diagnostic reports and thereby impeding large-scale precision medicine initiatives [8]. Specialized tools address specific needs: InterVar [9] enables semi-automated ACMG/AMP classification; ANNOVAR [10] extracts key variant features; PubTator 3.0 [11,12,13] annotates literature (see Fig. 1(b)); GPDminer [14] identifies variant-disease links. Integrated platforms like REVEL [18] (with MutPred [19,20] and FATHMM [21]) assess missense variants, whereas AutoPVS1 [22] automates high-throughput PVS1 evaluations. Despite their utility, these fragmented tools lack system-level interoperability to comprehensively extract insights from heterogeneous literature for clinical curation.

To address challenges in biomedical literature knowledge management, we provide a multimodal-driven, configurable platform for literature knowledge mining, delivering an end-to-end workflow from acquisition to extraction and integrating features surpassing existing systems (see Fig. 1(c)). Our contributions include two frameworks: 1) the Extensible Collection-Parsing-Structuring framework (*ECPS-LitTree*), which leverages an enhanced LayoutParser model [24] for multimodal parsing that converts PDFs into semi-structured JSON with text and image substructures. Unlike traditional methods [23], it automates literature collection through integration with PubMed Central and other databases while supporting user-uploaded PDFs, and introduces semantic-guided image segmentation and text extraction optimization for hierarchical storage of document elements via a literature tree architecture; and 2) the Configurable Requirement Customization (*CRC*) framework, which unites three established techniques in named entity recognition research: knowledge base-driven enumeration matching (e.g., ICD-10, HG38), regular expression-based extraction, and context-aware deep learning models. All benefiting from the continuous development of NER technology [29,30,31,32,33]. The *CRC* framework contributes a novel template mechanism that combines user-defined regular expressions with database entities to dynamically specify entity types (e.g., genes, diseases) and granular attributes, thereby overcoming the rigidity of conventional biomedical NER systems and enhancing complex semantic pattern recognition.

2 Methods

2.1 Design and implementation

The proposed biomedical literature mining platform integrates four modules: *ECPS-LitTree* for automated collection and multimodal parsing, *CRC* for dynamic entity filtering, *Data Engine* executing knowledge discovery via hybrid pattern mining (enumeration/rule-based/DL), and *Interactive Visualizer* for multidimensional evidence presentation. The adaptive workflow follows:

$$I = \begin{cases} V\Big(E\big(R(P(\mathbb{D}))\big)\Big), & \text{if } \mathbb{D} \in \mathcal{U} \text{ (user-uploaded)} \\ V\Big(E\big(R(P(C(\mathbb{D})))\big)\Big), & \text{if } \mathbb{D} \in \mathcal{S} \text{ (database-sourced)} \end{cases} \tag{1}$$

where: $\mathbb{D}$ denotes multi-source heterogeneous literature input comprising $\mathcal{U}$ (user-uploaded PDF documents) and $\mathcal{S}$ (database-sourced literature, e.g., PubMed Central); $C(\cdot)$ represents the Paper Collector for automated database acquisition when $\mathbb{D} \in \mathcal{S}$; $P(\cdot)$ indicates the Parser generating structured Literature Tree T; $R(\cdot)$ corresponds to the *CRC* Framework producing customized tree T'; $E(\cdot)$ signifies the Extractor mining information from T'; and $V(\cdot)$ embodies the Visualizer presenting final results through interactive dashboards.

2.2 *ECPS-LitTree* Framework

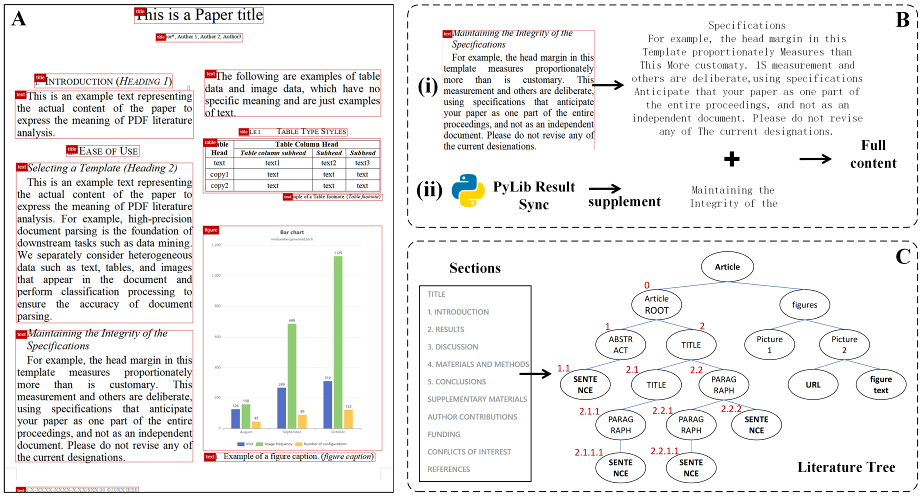
The *ECPS-LitTree* framework establishes a systematic pipeline for automated literature acquisition and multimodal parsing through open biomedical database integration. The workflow initiates dual data ingestion pathways: direct processing of user-uploaded PDFs or programmatic retrieval from integrated open-access repositories including PubMed Central, Europe PMC, and institutional repositories, formalized as:

$$\mathbb{D}_c = C(\mathbb{D}) \quad (\mathbb{D}\text{'s format is PDF}) \tag{2}$$

This multi-source strategy extends beyond conventional single-database scraping methods [23], expanding coverage to 94.6% of open-source biomedical literature types (CC-BY/NIH compliant) validated through NIH resource audits. The adaptive parsing function:

$$P(\mathbb{D}_c) = \text{LayoutParser}\left(S_{\text{text}}(R_{x\times y}) + S_{\text{image}}(R_{x\times y})\right) \tag{3}$$

where LayoutParser denotes the LayoutParser model [24], S_{text} and S_{image} represent processes that involve annotations and content extraction on the text and image parts of the PDF interface within the region $R_{x\times y}$, respectively, and $R_{x\times y}$ specifies the PDF page region, with x being the width and y being the height of the region. dynamically balances semantic confidence through multi-objective optimization, the Parser's annotation effect is demonstrated in Fig. 2A.



This is a Paper title

or*, Author 1, Author 2, Author3

I. INTRODUCTION (*HEADING 1*)

This is an example text representing the actual content of the paper to express the meaning of PDF literature analysis.

II. EASE OF USE

Selecting a Template (Heading 2)

This is an example text representing the actual content of the paper to express the meaning of PDF literature analysis. For example, high-precision document parsing is the foundation of downstream tasks such as data mining. We separately consider heterogeneous data such as text, tables, and images that appear in the document and perform classification processing to ensure the accuracy of document parsing.

Maintaining the Integrity of the Specifications

For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

X-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The following are examples of table data and image data, which have no specific meaning and are just examples of text.

LE I. TABLE TYPE STYLES

ble Head	Table Column Head		
	Table column subhead	*Subhead*	*Subhead*
text	text1	text2	text3
copy1	text	text	text
copy2	text	text	text

mple of a Table footnote. (*Table footnote*)

1. Example of a figure caption. (*figure caption*)

Maintaining the Integrity of the Specifications

For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Specifications
For example, the head margin in this
Template proportionately Measures than
This More customaty. 1S measurement and
others are deliberate,using specifications
Anticipate that your paper as one part of
the entire proceedings, and not as an
independent document. Please do not revise
any of The current designations.

Maintaining the
Integrity of the

Fig. 2. Schematic Diagram of the Literature-Parser. A: Schematic representation of document page segmentation annotation frames; B: Using mature Python parsing libraries to cross-reference and complement the results of text recognition for each section; C: "Literature Tree" Structure.

Our content extraction system integrates vision-based Detectron2 segmentation (domain-adapted weights) with Python multi-format parsing (PDF/XML/HTML fallback) (see Fig. 2B(i-ii)), achieving $O(n^{1.5})$ time complexity for n document elements. This dual-channel approach ensures cross-format stability through complementary graphical detection and raw text recovery mechanisms.

The enhanced Layout-Parser builds domain-specific literature trees T with structural constraints: $\mathit{Parent}(\text{ROOT}) = \{\text{Title} \prec \text{Abstract} \prec \text{Sentence}\}$ for text hierarchy, and $\mathit{Parent}(\text{Figures}) = \{\text{Picture} \prec (\text{URL} \oplus \text{FigText})\}$ for visual assets, where $\prec$ denotes containment and $\oplus$ parallel attributes (see Fig. 2C). The resultant literature tree T integrates these components through lattice-based fusion:

$$T = P(\mathbb{D}_c) + \{\mathit{Parent}(\text{ROOT}), \mathit{Parent}(\text{Figures})\} \tag{4}$$

where $+$ denotes the combination of literature content and structure. This multimodal integration generating FAIR-compliant JSON outputs that support federated literature mining.

2.3 *CRC* Framework

The *CRC* framework enables customizable biomedical entity definition and semantic filtering through template-driven configuration, formalized as:

$$E_{\text{target}} = \begin{cases} \bigcup_{k=1}^{K} (\text{Regex}_k \cap \text{Ontology}_k) & \text{(user-configured)} \\ \bigcup_{c \in \mathcal{C}_{\text{default}}} (\text{Regex}_c \cap \text{Ontology}_c) & \text{(default)} \end{cases} \tag{5}$$

where K denotes the number of user-defined regex-ontology pairs, with Regex_k representing pattern matching rules and Ontology_k specifying biomedical vocabularies (e.g., HG38 for genomes, ICD-10 for diseases). The default set $\mathcal{C}_{\text{default}}$ contains 10 atomic small templates: Gene, Transcript Number, Reference Genome Version, Variant, Family Chart Information, Experimental Result, Experimental Information, Disease, Species, and Phenotypic Data. Each small template combines:

$$\mathcal{T}_{\text{small}}^{i} := \underbrace{\text{Regex}_i \cap \text{Ontology}_i}_{\substack{\text{Core Logic} \\ \text{(syntax + semantics)}}} \times \underbrace{\text{Color}_i \times \text{Notes}_i}_{\substack{\text{Annotation} \\ \text{(custom fields)}}} \tag{6}$$

where $\cap$ requires dual pattern-ontology validation, and $\times$ denotes independent parameter dimensions. The configured literature tree is generated by:

$$T' = g_{\text{config}}(T, \mathcal{T}_{\text{large}}) \quad \text{where} \quad \mathcal{T}_{\text{large}} = \bigoplus_{m=1}^{M} \mathcal{T}_{\text{small}}^{s_m} \tag{7}$$

combines small templates through editable $\oplus$ operations that enable: (1) ontology-constrained pattern matching using structured vocabularies, and (2) free-form regex annotation with flexible text patterns. The $\oplus$ operator ensures valid template composition while preserving annotation consistency across combined elements. The semantic filtering mechanism employs a structure-aware relevance scoring approach that integrates literature tree T through:

$$\mathit{Relevance}(e,T) = \mathrm{Softmax}\left(\frac{Q_e \cdot \phi(T)^\top}{\sqrt{d}}\right) \tag{8}$$

where $\phi(T) = \mathrm{GAT}(T)$ encodes the literature tree using graph attention networks, $Q_e = \mathrm{MLP}(\mathrm{BERT}(e))$ combines entity embeddings with contextual features, d is the dimension of $\phi(T)$, and the superscript $\top$ denotes matrix transposition (distinct from the literature tree T). Entities are retained if $\mathit{Relevance}(e,T) > \theta = 0.4$ and $e \in \mathrm{ValidPath}(T)$, where $\mathrm{ValidPath}(T)$ denotes semantically coherent paths in the literature tree.

2.4 Data Mining

The data mining module extracts evidence through three synergistic strategies:

$$E(T') = E_{\mathrm{enum}}(T') \cup E_{\mathrm{rule}}(T') \cup E_{\mathrm{dl}}(T') \tag{9}$$

The multi-pattern matching framework combines the Enumeration Recognition and Regular Expressions, where $E_{\mathrm{enum}}(T') = \mathrm{AhoCorasick}(T', \{\mathrm{CHPO, HG38}\})$ achieves precise entity matching with $O(n+m)$ time complexity for n nodes and m patterns, while non-enumerable entities are processed through $E_{\mathrm{rule}}(T') = \bigcup \mathrm{RegExMatch}(T', \{\text{Regex Patterns}\})$. For deep learning recognition, the neural pipeline $E_{\mathrm{dl}}(T') = \mathrm{BioBERT}(T') \parallel \mathrm{BERN2}(T') \parallel \mathrm{AIONER}(T')$ implements parallel execution, where the operator $\parallel$ denotes independent model [4,25,26] execution to prevent prediction interference during joint inference.

2.5 Visualization of Results

The visualization module transforms structured literature data T' and mining results $E(T')$ into interactive reports through:

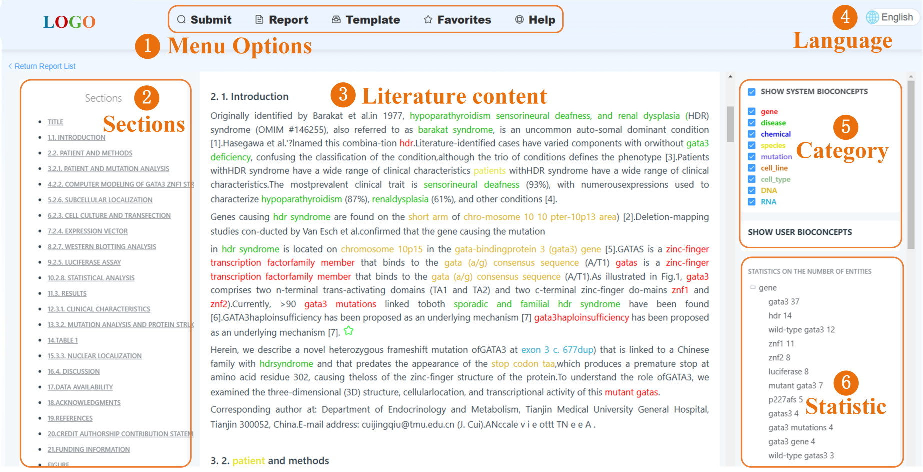
$$I = V(T', E(T')) = h_{\mathrm{vis}}(T', E(T')) \tag{10}$$

where $h_{\mathrm{vis}}(\cdot)$ implements a dual-pane interface (see Fig. 3): Left Pane displays categorical entity statistics (document navigation) and Right Pane provides interactive filters (entity-type toggles, frequency distributions, evidence metrics). Additional features include bookmarking (paragraph/sentence archival) and Excel export of filtered reports.

3 Experiments and Results

3.1 Experimental Setup

Our experiments implemented $\theta = 0.4$ (chosen as an optimal balance between precision and recall within the $0.3 - 0.7$ range) *relevance* threshold filtering ($\mathit{Relevance}(e) > 0.4$) with the AIONER-based framework $E_{\mathrm{dl}}(T') = \mathrm{AIONER}(T')$ for biomedical entity recognition. Data combined open-source repositories and user-uploaded PDFs, with controlled variation across trials using distinct PMC-OA and institutional literature corpora to validate methodological robustness.



- TITLE
- 1.1. INTRODUCTION
- 2.2. PATIENT AND METHODS
- 3.2.1. PATIENT AND MUTATION ANALYSIS
- 4.2.2. COMPUTER MODELING OF GATA3 ZNF1 STR
- 5.2.6. SUBCELLULAR LOCALIZATION
- 6.2.3. CELL CULTURE AND TRANSFECTION
- 7.2.4. EXPRESSION VECTOR
- 8.2.7. WESTERN BLOTTING ANALYSIS
- 9.2.5. LUCIFERASE ASSAY
- 10.2.8. STATISTICAL ANALYSIS
- 11.3. RESULTS
- 12.3.1. CLINICAL CHARACTERISTICS
- 13.3.2. MUTATION ANALYSIS AND PROTEIN STRU
- 14.TABLE 1
- 15.3.3. NUCLEAR LOCALIZATION
- 16.4. DISCUSSION
- 17.DATA AVAILABILITY
- 18.ACKNOWLEDGMENTS
- 19.REFERENCES
- 20.CREDIT AUTHORSHIP CONTRIBUTION STATEM
- 21.FUNDING INFORMATION
- FIGURE

2. 1. Introduction

Originally identified by Barakat et al.in 1977, hypoparathyroidism sensorineural deafness, and renal dysplasia (HDR) syndrome (OMIM #146255), also referred to as barakat syndrome, is an uncommon auto-somal dominant condition [1].Hasegawa et al.'?Inamed this combina-tion hdr.Literature-identified cases have varied components with orwithout gata3 deficiency, confusing the classification of the condition,although the trio of conditions defines the phenotype [3].Patients withHDR syndrome have a wide range of clinical characteristics patients withHDR syndrome have a wide range of clinical characteristics.The mostprevalent clinical trait is sensorineural deafness (93%), with numerousexpressions used to characterize hypoparathyroidism (87%), renaldysplasia (61%), and other conditions [4].

Genes causing hdr syndrome are found on the short arm of chro-mosome 10 10 pter-10p13 area) [2].Deletion-mapping studies con-ducted by Van Esch et al.confirmed that the gene causing the mutation

in hdr syndrome is located on chromosome 10p15 in the gata-bindingprotein 3 (gata3) gene [5].GATAS is a zinc-finger transcription factorfamily member that binds to the gata (a/g) consensus sequence (A/T1) gatas is a zinc-finger transcription factorfamily member that binds to the gata (a/g) consensus sequence (A/T1).As illustrated in Fig.1, gata3 comprises two n-terminal trans-activating domains (TA1 and TA2) and two c-terminal zinc-finger do-mains znf1 and znf2).Currently, >90 gata3 mutations linked toboth sporadic and familial hdr syndrome have been found [6].GATA3haploinsufficiency has been proposed as an underlying mechanism [7] gata3haploinsufficiency has been proposed as an underlying mechanism [7]. ☆

Herein, we describe a novel heterozygous frameshift mutation ofGATA3 at exon 3 c. 677dup) that is linked to a Chinese family with hdrsyndrome and that predates the appearance of the stop codon taa,which produces a premature stop at amino acid residue 302, causing theloss of the zinc-finger structure of the protein.To understand the role ofGATA3, we examined the three-dimensional (3D) structure, cellularlocation, and transcriptional activity of this mutant gatas.

Corresponding author at: Department of Endocrinology and Metabolism, Tianjin Medical University General Hospital, Tianjin 300052, China.E-mail address: cuijingqiu@tmu.edu.cn (J. Cui).ANccale v i e ottt TN e e A .

3. 2. patient and methods

☑ SHOW SYSTEM BIOCONCEPTS

☑ gene
☑ disease
☑ chemical
☑ species
☑ mutation
☑ cell_line
☑ cell_type
☑ DNA
☑ RNA

SHOW USER BIOCONCEPTS

STATISTICS ON THE NUMBER OF ENTITIES

gene
gata3 37
hdr 14
wild-type gata3 12
znf1 11
znf2 8
luciferase 8
mutant gata3 7
p227afs 5
gatas3 4
gata3 mutations 4
gata3 gene 4
wild-type gatas3 3

Fig. 3. Screenshot of the Report Viewing Interface: (1) Menu Options; (2) Article Sections; (3) Literature content Annotation Effects; (4) Language Switching; (5) Annotation Options of Category; (6) Entity Information Statistics.

3.2 Functional Comparison and Identification Accuracy

Table 1 summarizes a feature comparison between our platform and other state-of-the-art text mining tools (PubTator3.0 [13], GPDminer [14], PubTerm [27], and PubMedKB [28]). Notably, our platform uniquely supports dynamic requirement configuration and advanced result visualization.

Table 1. Function comparison with text mining tools.

Tool	NER	Literature Upload	Visualization of the original paper	Requirement configuration
Our Platform	✓	✓	✓	✓
PubTator3.0 [13]	✓	—	✓	—
GPDminer [14]	✓	—	—	—
PubTerm [27]	✓	—	—	—
PubMedKB [28]	✓	—	✓	—

3.3 Report Generation Stability

We evaluated report generation stability through three controlled experiments using different submission methods: direct PMID submission, PMID submission via Excel, and direct PDF submission. Overall success rates were 87%, 82%, and 100% respectively. The average report generation times (including all processing stages) were 3.72 minutes per literature for PMID, 3.65 minutes for PMID via

Excel, and 3.35 minutes for PDF submission (see Fig. 4A). These results confirm that, while all methods are efficient, direct PDF submission is optimal due to bypassing the literature collection stage.

3.4 Paper Collection and Data Mining Volume Case Study

A large-scale simulation using 800 biomedical articles (retrieved via PMID) demonstrated the superior performance of our platform over PubTator 3.0. As shown in Fig. 4B, our platform acquired 758 full-text articles (with 42 articles yielding no results), achieving a full-text acquisition rate of 94.75% compared to PubTator 3.0's 42% (with the majority as abstracts).And Fig. 4C(i) and C(ii) Comparison of different types of information mining volume under the same condition. This comprehensive evaluation demonstrates that the platform's literature acquisition capability, knowledge extraction capacity, and reliability are positioned at the forefront of the field.

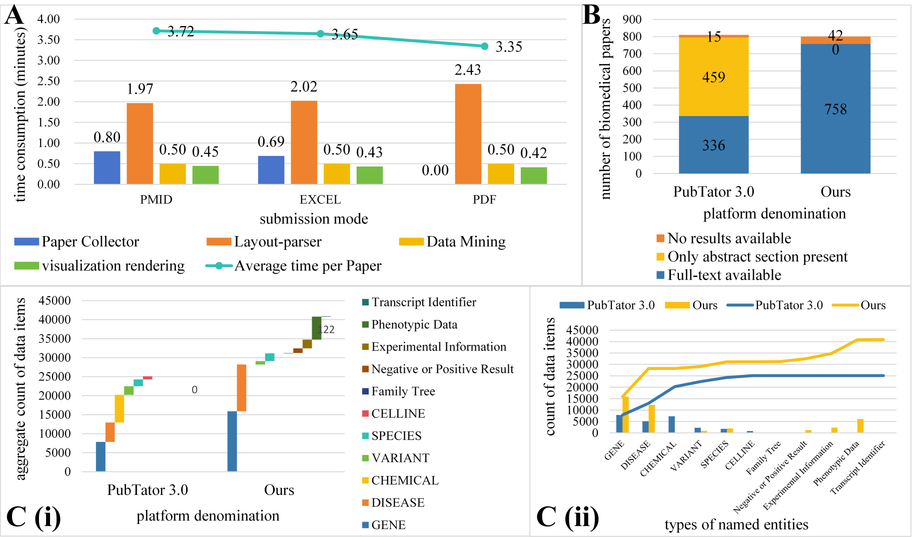


Fig. 4. (A) Average time per literature for each stage and overall; (B) Article retrieval and collection comparison between our platform and PubTator 3.0; (C-i) Overall data mining volume comparison; (C-ii) Detailed category-wise data mining volume.

4 Conclusion

This study presents a robust biomedical literature mining platform integrating multimodal retrieval, adaptive parsing, configurable entity recognition, and interactive visualization. Experimental evaluations demonstrate superior performance in complex entity extraction tasks, outperforming existing tools through enhanced configurability, operational flexibility, and notable efficiency gains. Future work will extend to semantic relationship mining while integrating systematic evaluation and application of Large Language Models (LLMs), aiming to further strengthen unstructured knowledge discovery within optimized text mining pipelines.

Acknowledgments. This research was funded by the General Program of the National Natural Science Foundation of Hunan Province, China (Grant Nos. 2025JJ70028, 2025JJ81178, 2024JJ9550, and 2023JJ30975), the Key Scientific Research Project of the Education Department of Hunan Province, China (Grant No. 24A0401), and the Postgraduate Scientific Research Innovation Project of Hunan Province, China (Grant No. LXBZZ2024316). We gratefully acknowledge the Bioinformatics Center of Furong Laboratory and Xiangya Hospital (Central South University) for their invaluable support in providing essential computational resources and server infrastructure.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dorsch, J.L., Faughnan, J.G., Humphreys, B.L.: Grateful Med: Direct access to MEDLINE for health professionals with personal computers. *Information Services and Use* **42**(2), 151–160 (2022)
2. De Paoli, F., et al.: VarChat: the generative AI assistant for the interpretation of human genomic variations. *Bioinformatics* **40**(4), btae183 (2024)
3. Pasche, E., et al.: Variomes: a high recall search engine to support the curation of genomic variants. *Bioinformatics* **38**(9), 2595–2601 (2022)
4. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)
5. Richards, S., et al.: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**(5), 405–424 (2015)
6. Wang, Z., et al.: VarCards2: an integrated genetic and clinical database for ACMG-AMP variant-interpretation guidelines in the human whole genome. *Nucleic Acids Research* **52**(D1), D1478–D1489 (2024)
7. Karpova, N., Dmitrenko, O., Arshinova, E.: In Silico Determination of Changes in Transcription Factor Binding Sites for the Preeclampsia Risk Haplotype in the Regulatory Region of the FLT1 Gene. In: *Proceedings of the 25th International Conference on Engineering and Computer-Based Medical Systems (IECBM 2022)*, pp. 1–31. MDPI, Basel (2022). https://doi.org/10.3390/iecbm2022-13721
8. Rajman, M., Besançon, R.: Text mining: natural language techniques and text mining applications. In: *Data Mining and Reverse Engineering: Searching for Semantics*. 7th IFIP Conference on Database Semantics (DS-7), Leysin, Switzerland, October 7–10, 1997, pp. 50–64. Springer, Heidelberg (1998)

9. Li, Q., Wang, K.: InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *The American Journal of Human Genetics* **100**(2), 267–280 (2017)
10. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**(16), e164–e164 (2010). https://doi.org/10.1093/nar/gkq603
11. Wei, C.-H., Kao, H.-Y., Lu, Z.: PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* **41**(W1), W518–W522 (2013)
12. Wei, C.-H., et al.: PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research* **47**(W1), W587–W593 (2019)
13. Wei, C.-H., et al.: PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research* **52**(W1), W540–W546 (2024)
14. Park, Y.-J., et al.: GPDminer: a tool for extracting named entities and analyzing relations in biological literature. *BMC Bioinformatics* **25**(1), 101–101 (2024)
15. Riggs, E.R., et al.: Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine* **22**(2), 245–257 (2020)
16. Stenson, P.D., et al.: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics* **133**(1), 1–9 (2014)
17. Amberger, J.S., Hamosh, A.: Searching Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic phenotypes. *Current Protocols in Bioinformatics* **58**, 1.2.1–1.2.12 (2017). https://doi.org/10.1002/cpbi.27
18. Ioannidis, N.M., et al.: REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**(4), 877–885 (2016)
19. Mort, M., et al.: MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology* **15**(1), R19–R19 (2014)
20. Shanmugam, N.R.S., Veluraja, K., Gromiha, M.M.: PCA-MutPred: prediction of binding free energy change upon missense mutation in protein-carbohydrate complexes. *Journal of Molecular Biology* **434**(11), 167526 (2022)
21. Rogers, M.F., et al.: FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**(3), 511–513 (2018)
22. Xiang, J., Peng, J., Baxter, S., Peng, Z.: AutoPVS1: an automatic classification tool for PVS1 interpretation of null variants. *Human Mutation* **41**(9), 1488–1498 (2020)
23. Comeau, D.C., et al.: PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics* **35**(18), 3533–3535 (2019)
24. Shen, Z., et al.: LayoutParser: a unified toolkit for deep learning-based document image analysis. In: *Document Analysis and Recognition – ICDAR 2021: 16th International Conference*, pp. 131–146. Springer, Lausanne (2021)
25. Sung, M., et al.: BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **38**(20), 4837–4839 (2022)
26. Luo, L., et al.: AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics* **39**(5), btad310 (2023)
27. Garcia-Pelaez, J., et al.: PubTerm: a web tool for organizing, annotating and curating genes, diseases, molecules and other concepts from PubMed records. *Database: The Journal of Biological Databases and Curation* **2019**, bay137–bay137 (2019)

28. Li, P.-H., et al.: PubMedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Research* **50**(W1), W616–W622 (2022). https://doi.org/10.1093/nar/gkac310
29. Chen, K., et al.: TRBNER: named entity recognition of TCM medical records based on multi-feature fusion. In: *IET Conference Proceedings CP989*, vol. 2024, no. 21, pp. 174–181. IET, London (2024)
30. Zhang, L., et al.: Chinese sequence labeling with semi-supervised boundary-aware language model pre-training. arXiv preprint arXiv:2404.05560 (2024).
31. Geroldi, A., et al.: Next-generation sequencing in Charcot-Marie-Tooth: a proposal for improvement of ACMG guidelines for variant evaluation. *Journal of Medical Genetics* **61**(9), 847–852 (2024). https://doi.org/10.1136/jmg-2024-110019
32. Goyal, A., et al.: Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review* **29**, 21–43 (2018)
33. Dogan, C., et al.: Fine-grained named entity recognition using ELMo and Wikidata. arXiv preprint arXiv:1904.10503 (2019)