Advances in Biomedical Literature Mining: A Comparative Review of Tools and the Emergence of the Swalife AI-Powered Platform

Dr. Pravin Badhe[1], Ashwini Badhe[2]

[1,2] Swalife Biotech Ltd, North Point House, North Point Business Park, Cork, Republic of Ireland

Corresponding author Email: drpravinbadhe@swalifebiotech.com
Doi: https://doi.org/10.5281/zenodo.17718658
Received: 16 March 2026
Accepted: 26 March 2026

Abstract

The biomedical literature continues to expand at an unprecedented rate, creating significant challenges for researchers to efficiently extract actionable knowledge. Traditional literature mining tools often focus on narrow domains or lack integration of natural product research, limiting their utility in drug discovery. Artificial intelligence (AI) and natural language processing (NLP) have revolutionized biomedical text mining, enabling higher accuracy and broader scope. This review critically evaluates major biomedical literature mining platforms, highlighting their strengths and limitations, especially concerning herbal therapeutics and protein target analysis. We introduce the Swalife AI-powered tool, which uniquely integrates herb–disease–protein mining with AI-driven filtering and dynamic network visualization. Swalife addresses critical gaps by delivering a comprehensive, exportable platform tailored to natural product research, expediting drug discovery and systems pharmacology investigations.

Keywords: Literature Mining, Text Mining, Bioinformatics Tools, Network Pharmacology, AI in Drug Discovery, Herbal Research

Introduction

The volume of biomedical publications indexed in databases such as PubMed has escalated dramatically, with over 35 million citations as of 2025 and daily additions exceeding thousands. This volume outpaces manual review capabilities, creating a bottleneck that impedes knowledge translation into therapeutic advances (Lu 2011). Automated literature mining approaches have become essential to systematically extract relationships among biological entities including genes, diseases, and chemical compounds (Wei et al., 2012). While early tools relied on keyword searches, AI-driven natural language processing has enabled nuanced extraction of interactions and biological context (Hopkins et al., 2015). Despite advances, many systems inadequately address multi-domain integration, particularly the complex relationship triads of herbs, diseases, and protein targets relevant to natural product drug discovery (Kell 2020). This review aims to evaluate notable literature mining tools, focusing on their capabilities in herbal and protein-centric contexts, and to highlight the novel contribution of the Swalife AI platform designed to overcome current limitations.

Traditional Approaches to Literature Mining before computational tools, researchers relied on labor-intensive manual literature review prone to bias and incompleteness (Zhao et al., 2019). Early computational tools applied keyword and Boolean searches but lacked semantic understanding and relationship inference capabilities. The advent of natural language processing and machine learning in the 2010s spurred tools capable of named entity recognition (NER), relationship extraction, and contextual relevance scoring (Cohen and Hersh, 2005). These

tools enabled automated identification of gene-disease and variant-phenotype relationships but often ignored natural product contexts. Recent works incorporate AI to enhance relevance and reduce noise but remain fragmented in scope and integration

Overview of Existing Literature-Mining Tools We reviewed key tools based on input flexibility, literature extraction, visualization, herb–disease–protein mapping, and AI sophistication:

- PubTator Central provides automated NER for genes and proteins across biomedical abstracts but lacks direct herbal or compound relationship mining (Wei et al., 2013).

- LitVar specializes in genomic variant interpretation but does not support herbal or multi-domain queries (Allot, A. et al., 2016).

- Textpresso focuses on model organisms with customizable corpora but is limited for multi-domain drug discovery contexts (Müller et al., 2018).

- iHOP mines protein–protein interactions at sentence level but has limited capabilities with herbal compounds (Hoffmann and Valencia, 2005).

- CoreMine Medical applies concept co-occurrence with basic visualization but provides limited compound prioritization and user customizability.

- DisGeNET, STRING, and GeneCards offer curated gene-disease associations and protein interaction maps but do not perform real-time literature mining or herbal data integration (Piñero et al., 2020; Szklarczyk et al., 2021).

Table 1: Comparative Analysis of Tools

Feature	**PubTator**	**LitVar**	**Textpresso**	**iHOP**	**CoreMine**	**DisGeNET /STRING**	**Swalife**
Herb/Natural Product Mining	No	No	No	No	No	No	Yes
Disease Filtering	No	Yes	Limited	No	Limited	Yes	Yes
Protein Target Extraction	Yes	No	No	Yes	Limited	Yes	Yes

Feature	PubTator	LitVar	Textpresso	iHOP	CoreMine	DisGeNET /STRING	Swalife
Visualization	No	No	No	No	Yes	Yes	Yes
Exportable Structured Data	No	No	No	No	Limited	Limited	Yes
AI/NLP Depth	Moderate	Moderate	Moderate	Limited	Limited	No	High
User Accessibility	High	Moderate	Moderate	Moderate	Moderate	Moderate	High

The table underscores Swalife's unique ability to mine herb–disease–protein relationships with AI-powered relevance filtering and interactive network visualization.

The Swalife AI-Powered Literature Mining Tool

Swalife is a web-based platform integrating AI-enhanced mining from PubMed abstracts combined with UniProt protein databases. It allows simultaneous queries on herbs and diseases, extracting relationships supported by underlying literature. Its network visualization module dynamically displays herb, disease, and protein nodes with weighted edges representing interaction strength and literature support. Structured datasets including protein lists, associated literature, and network files are exportable, facilitating downstream analyses such as network pharmacology and docking simulations. This integrated workflow markedly reduces research time and enhances multi-target hypothesis generation (Badhe 2025).

How Swalife Outperforms Existing Tools

Swalife's specialization in herbal therapeutics plus disease and protein targets fills critical gaps not addressed by gene-centric or disease-centric tools. Its integrated pipeline streamlines mining, filtering, protein mapping, and visualization in one interface, eliminating the need for multiple platforms. AI-driven relevance algorithms prune irrelevant or redundant articles, enhancing result quality. The unique herb–disease–protein network architecture delivers molecular insights central to natural product drug discovery. Exportable, structured data empower bioinformaticians to employ Swalife outputs in multi-omics and docking pipelines, accelerating translational research.

Limitations and Future Scope of Swalife

Despite its innovative design, the Swalife AI-powered literature mining tool has several important limitations. First, the platform's reliance on public biomedical databases such as PubMed and UniProt inherently restricts the

scope of extractable data. Novel herbs or proteins that have limited or no representation in these databases will be underrepresented or entirely omitted. This constrains the tool's ability to comprehensively cover emerging discoveries or less-studied natural compounds.

Second, Swalife currently mines only abstracts and article metadata rather than full-text articles. Abstract mining, while efficient, may miss detailed mechanistic insights, nuanced protein interactions, or herbal pharmacodynamics often described in full texts. This restricts the depth of biological inference and may result in partial or incomplete relationship extraction.

Third, the relationships identified by Swalife are computationally predicted using AI algorithms, which, while robust, require experimental validation. The tool serves as a hypothesis-generation platform rather than a definitive source of confirmed biological interactions. Consequently, users are advised to interpret results as leads for further wet-lab testing.

Fourth, entity recognition-inclusive of herbs, diseases, and proteins-is susceptible to ambiguity, synonymy, and polysemy inherent in biomedical nomenclature. Although the AI reduces errors, some spurious or missed associations remain inevitable, especially for entities with overlapping or non-standard terminology.

Looking ahead, Swalife's roadmap includes several promising extensions to strengthen its capabilities. Integrating additional databases beyond PubMed and UniProt, such as Scopus, Web of Science, or specialized phytochemical repositories, would improve coverage and data diversity. The planned incorporation of molecular docking pipelines aims to bridge literature-derived hypotheses with structural and chemical interaction validation, thereby increasing biological relevance. Multi-omics data layers-including genomics, transcriptomics, and metabolomics-will enhance the network's depth and enable more systems-level analyses. Collaborative annotation features could harness community expert input to refine entity recognition and relationship accuracy over time. Lastly, implementing full-text mining and natural language processing of entire articles will enable richer mechanistic insights and comprehensive data capture.

Conclusion

The overwhelming growth of biomedical literature demands automated mining tools tailored to domain-specific needs. While many existing platforms provide valuable functionality, they primarily target genomic or protein-centric analyses and lack integration of herbal or natural product contexts essential to modern phytopharmacology. In this landscape, the Swalife AI-powered platform offers a pioneering, comprehensive herb–disease–protein mining ecosystem, uniquely combining AI-driven literature mining with protein database integration and advanced network visualization. By bridging traditional herbal knowledge with contemporary molecular data, it significantly accelerates drug discovery workflows, systems pharmacology research, and personalized medicine development. Its unique focus on natural products positions it as an indispensable tool in phytochemical research, filling a critical gap in biomedical informatics. As enhancements in database breadth, AI algorithms, docking integration, and multi-omics layering unfold, the tool is poised to become a foundational platform advancing integrative biomedical research and therapeutic innovation.

References

1. Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, *2011*, baq036.

2. Wei, C. H., Harris, B. R., Kao, H. Y., & Lu, Z. (2013). tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, *29*(11), 1433-1439.

3. Wei, C. H., Phan, L., Feltz, J., Maiti, R., Hefferon, T., & Lu, Z. (2018). tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*, *34*(1), 80-87.

4. Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, *4*(11), 682-690.

5. Zhao, S., Su, C., Lu, Z., & Wang, F. (2021). Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, *22*(3), bbaa057.

6. Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics*, *6*(1), 57-71.

7. Wei, C. H., Kao, H. Y., & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, *41*(W1), W518-W522.

8. Allot, A., Peng, Y., Wei, C. H., Lee, K., Phan, L., & Lu, Z. (2018). LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic acids research*, *46*(W1), W530-W536.

9. Müller, H. M., Van Auken, K. M., Li, Y., & Sternberg, P. W. (2018). Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC bioinformatics*, *19*(1), 94.

10. Hoffmann, R., & Valencia, A. (2004). A gene network for navigating the literature. *Nature genetics*, *36*(7), 664-664.

11. Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, *48*(D1), D845-D855.

12. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... & Mering, C. V. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), D607-D613.

13. Badhe, P. (2025). Swalife AI-Powered Network Analysis Tool: An Intelligent Platform for Herb-Disease-Protein Interaction Mining.