BioTopicXplor: A Web Tool for Interactive Exploration of PubMed Literature through Reproducible Topics.

Lana Yeganova lana.yeganova@nih.gov
Donald C. Comeau donald.comeau@nih.gov
Won Kim wonkim@nih.gov
Natalie Xie natxie@nih.gov
Shubo Tian shubo.tian@nih.gov
W. John Wilbur john.wilbur@nih.gov
Zhiyong Lu zhiyong.lu@nih.gov

Division of Intramural Research (DIR), NLM, NIH, Bethesda, MD USA 20894

Abstract

The rapid expansion of biomedical literature presents a major challenge for researchers seeking to organize knowledge, identify emerging scientific trends, and explore large PubMed result sets. Existing literature exploration systems primarily rely on predefined ontologies, user-selected features, or clustering approaches that may produce unstable topic structures across runs. We present BioTopicXplor, a tool for interactive exploration of user-defined PubMed article collections through reproducible topic discovery and hierarchical organization.

BioTopicXplor is powered by ConvexTopics, an optimization-based topic modeling framework that formulates topic discovery as a convex optimization problem with mixture model parameters drawn directly from the data. Unlike traditional clustering approaches that depend on random initialization and may converge to local optima, ConvexTopics guarantees convergence to a global optimum, eliminates the need to predefine the number of topics, and produces stable, fine-grained, and interpretable topic structures across repeated runs on the same corpus.

To support intuitive exploration, BioTopicXplor integrates large language models to generate concise literature-grounded summaries and organize fine-grained subtopics into higher-level thematic groups. The system combines reproducible topic modeling with an interactive human-centric interface that enables users to navigate biomedical literature collections from broad conceptual themes to specific subtopics and associated PubMed articles. BioTopicXplor is freely available at: https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/topics/dev/create.html.

1 Introduction

The exponential growth of biomedical publications presents a critical challenge for researchers seeking to organize, interpret, and discover knowledge or emerging scientific trends. PubMed (https://pubmed.ncbi.nlm.nih.gov/) remains the main gateway to biomedical literature and provides access to relevant documents given a query. However, it does not offer support for understanding the conceptual structure of a collection.

Several companion tools enable alternative options for exploring biomedical literature. Anne O'Tate (Engwall, 2017) enables users to analyze PubMed search results by ranking and grouping articles based on important words in titles and abstracts. It also allows clustering by MeSH term–based topics, author names, affiliations, journal names, and publication year. BioTextQuest v2.0 (Theodosiou et al., 2024), extends this paradigm by offering interactive clustering based on user-selected biomedical terms. Ontology-based interfaces such as GoPubMed (Doms and Schroeder, 2005) organize PubMed records within structured vocabularies, including Gene Ontology and medical Subject Headings (MeSH), facilitating exploration but remaining constrained to predefined ontological categories rather than enabling unsupervised topic discovery. While these tools facilitate exploration, they rely on predefined features and user-selected terms, or curated ontologies to organize the literature. As a result, they may be limited in their ability to capture the data-driven structure of a document collection or to reveal fine-grained, emergent topics in an unsupervised way.

To address these limitations, we present BioTopicXplor, a tool for interactive exploration of user-defined PubMed article collections. Given a PubMed query, BioTopicXplor performs on-demand clustering/topic discovery and organizes the resulting topics into an intuitive representation that supports exploration of literature collections both through navigating the topics and querying within the topics.

BioTopicXplor is powered by ConvexTopics

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 475–480

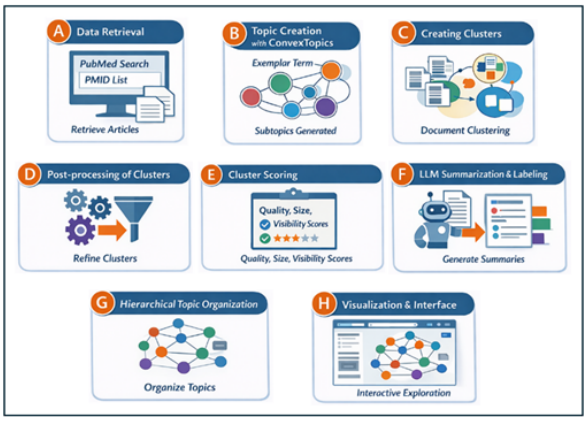

Figure 1: BioTopicXplor architecture.

(Yeganova et al., 2026), an optimization-based topic modeling framework that is formulated as a convex optimization problem with exemplars drawn directly from the data (Lashkari and Golland, 2007). We chose this approach to avoid limitations of traditional clustering methods such as K-means (Ja, 1979) or expectation–maximization for mixture models (Blei et al., 2003). These suffer from sensitivity to initialization and convergence to local optima, limiting reproducibility and scalability on large corpora. As a result, these methods generally produce different document or term groupings across runs.

ConvexTopics guarantees convergence to a global optimum, which means it yields the same set of topics across repeated runs on the same document collection. Additionally, it eliminates dependence on random initialization, and the need to predefine the number of topics. The algorithm produces a large number of fine-grained topics, referred to as subtopics.

To support interpretation and exploration of these subtopics, BioTopicXplor integrates large language models for two key tasks: (1) to generate concise summaries for each subtopic, and (2) to construct a second layer that organizes subtopics into higher-level groups, referred to as major topics. This hierarchical structure enables users to navigate from broad thematic areas to more specific ones. In addition, the tool provides a search functionality that allows users to query and explore subtopics interactively. We demonstrate the utility of the system through application to anti-aging literature, illustrating its ability to reveal meaningful thematic structure leading to knowledge discovery.

2 System Overview

Given a user-specified PubMed query, BioTopicXplor retrieves the corresponding articles, applies the ConvexTopics framework to generate a set of subtopics, and subsequently organizes them into higher-level, more general topics. The overall pipeline (illustrated in Figure 1) integrates data retrieval, text processing, optimization-based topic modeling, document clustering, and visualization components, as well as large language models to generate summaries for each subtopic, allowing users to seamlessly move from raw literature collections to structured, interpretable topics.

As LLMs become increasingly integrated into biomedical text mining, their ability to generate user-friendly summaries is highly valuable. However, they remain prone to hallucination. ConvexTopics provides a stable backbone that constrains the use of LLMs, ensuring that summaries are grounded in algorithmically derived topics. We believe this controlled integration provides the benefit of human-readable summaries while minimizing hallucinations by keeping the summaries tightly anchored to the topic terms and relevant documents computed by ConvexTopics.

Data Retrieval. BioTopicXplor begins by retrieving a set of PubMed articles based on a user-specified query. By default, the retrievals are limited to the last five years. The system collects titles and abstracts for downstream processing.

Creating Subtopics with ConvexTopics The subtopics are computed using the ConvexTopics algorithm (Yeganova et al., 2026), which reformulates clustering as a convex optimization problem by restricting potential cluster centers to exemplars drawn directly from the data. This transforms what is ordinarily a combinatorial search into a tractable convex objective that is guaranteed to converge to a globally optimal solution (Lashkari and Golland, 2007). Applied to biomedical text, the algorithm operates over a vocabulary of statistically enriched single terms and noun phrases, as many biomedical concepts are expressed as multi-word expressions (Yeganova et al., 2009, 2011). Term similarity is measured via the Dice coefficient (Manning, 2008). The number of resulting subtopics is determined automatically during computation, yielding fine-grained, reproducible topic structures without manual tuning. Each subtopic is defined by an exemplar and a set of representative terms that are used to rank documents by relevance.

Creating Document Clusters. For each subtopic created above, we use its topic terms to compute the list of the most relevant PubMed articles. These documents are selected based on their relevance to the topic terms and provide a cluster of documents associated with each topic. Clicking on a PubMed ID or article title leads to the article view in PubMed, where one can read, download, or follow links to full-text, if available.

Post-processing of Clusters. Following initial clustering, a post-processing step is applied. Redundant clusters and clusters with substantial overlap (greater than 50% shared documents) are removed to ensure a diverse and non-redundant set of subtopics.

Cluster Scoring. Each topic/cluster is assigned three scores based on the following criteria: (i) an intrinsic cluster quality score, (ii) the size of the cluster measured by the number of associated documents, and (iii) a visibility score that reflects the prominence of the ten top scoring topic terms in the ten top scoring documents. Each of these scores can be used to rank subtopics for presentation.

Summarization and Labeling. BioTopicXplor employs large language models to generate concise, literature-grounded summaries and descriptive titles for each subtopic. The summaries are generated based on the top ten documents associated with each cluster using GPT-5.1. Each generated summary sentence is linked to PMID(s) that were used for extracting the underlying evidence. The title is generated to best describe the summary.

Hierarchical Topic Organization. To organize the large number of subtopics for presentation, they are grouped into higher-level topics (major topics). This step leverages large language models to group semantically related subtopics and construct a hierarchical representation.

Visualization and Interface. The resulting higher-level topics are presented through an interactive interface that allows users to explore clusters, view summaries, and access associated PubMed articles. The visualization is designed to provide an overview of the conceptual structure of literature.

Implementation Details. The user-facing web interface was developed in-house using standard HTML, CSS, and JavaScript. The front end is compatible with modern desktop browsers and iPads. Although not specifically optimized for phone screens, the server is functional on mobile devices. User requests are submitted to the server, which was also developed in-house using C++ and Python. Both the front end and backend components run on a Linux-based system. The system runs on a shared machine equipped with a 12-core Xeon(R) Gold CPU operating at up to 2.30 GHz and 196 GB of RAM. Most of the processing time is spent generating cluster titles and summaries using an OpenAI GPT model, currently GPT-5.1.

3 Usage

Users initiate an analysis by submitting a PubMed query, after which the system automatically retrieves the corresponding articles and executes the full processing pipeline. The results are provided through a uniquely generated web link. Multiple analyses can be conducted concurrently, and previously generated instances remain available for subsequent exploration.

View Major Topics. On the main page, users are presented with Major Topics derived for the dataset. These serve as starting points for exploration. Selecting a Major Topic leads to corresponding Subtopics, which can be ranked according to multiple scoring criteria.

Explore Subtopics. Each subtopic is associated with a dedicated page that includes a concise summary and title generated from the ten most relevant documents, a ranked list of associated PubMed articles with links for direct access, and a set of topic terms defining the topic.

Explore the data through Search. In addition to hierarchical navigation, the system supports keyword-based search across topics. Search can be performed globally across the entire collection or restricted to a specific major topic, enabling targeted exploration of relevant subtopics.

Use Case: Exploring Anti-Aging literature in PubMed. The growing interest in longevity and anti-aging has led to a proliferation of resources promoting various interventions, not all of which are supported by rigorous scientific evidence. This underscores the need for systematic analysis of biomedical literature to distinguish evidence-based claims. Using anti-aging research as a representative use case, (Yeganova et al., 2026) demonstrates how BioTopicXplor, applied to 12,000 PubMed articles, captures the research landscape through nuanced subtopics with summaries, and hierarchical organization.

Figure 2 illustrates the framework. The left panel shows the main page, where users are presented with a list of major topics. The right panel shows

a detailed topic page. Demonstrated in Figure 2 is a topic page produced by choosing the "Gut Microbiota and aging" major topic and selecting the "Beneficial Bacteria and Natural Compounds Combat Aging" subtopic.

Using this analysis, (Yeganova et al., 2026) observed that although NAD supplements are widely available, their safety remains insufficiently supported by evidence. Key uncertainties include pharmacokinetics and pharmacodynamics, particularly bioavailability, metabolism, and tissue specificity (Poljšak et al., 2022). Moreover, the lack of long-term safety studies and limited clinical trials leaves optimal dosing unclear, with some studies suggesting a potential risk of kidney inflammation in older populations (Song et al., 2023; Nadeeshani et al., 2022).

4 Evaluation

Many factors can be considered for evaluating BioTopicXplor as an end-to-end system, examining not only the quality of the topics produced by ConvexTopics but also the interpretability and utility of the resulting framework.

The quality of Subtopics. The ConvexTopics algorithm has been evaluated on both general-purpose NLP benchmarks and biomedical text corpora in (Yeganova et al., 2026). The annotated baselines include 20-Newsgroups (`http://people.csail.mit.edu/jrennie/20Newsgroups`) and Reuters-21578 (Lewis et al., 2004), while the biomedical datasets consist of PubMed articles on Anti-Aging, Diabetes Mellitus, and Age-Related Macular Degeneration. ConvexTopics was compared with state-of-the-art clustering and topic modeling approaches, including K-means (Ja, 1979; Lloyd, 1982), LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022) and was shown to perform favorably.

The quality of Major Topics. To facilitate navigation, BioTopicXplor organizes subtopics into major topics using large language models. We evaluated this hierarchical structure through manual inspection. The major topics were observed to correspond to meaningful thematic groupings, providing a coherent high-level view of the literature while preserving access to detailed subtopics. While some variations in major topics may occur accross the runs, the resulting groupings remain meaningful.

Timing. Depending on the size of the dataset, processing may take up to several hours. For example, analysis of 1,099 documents retrieved using the query “familial Mediterranean fever” (limited to 5 most recent years) completed in 65 minutes, while the same query without time restriction produced 5,731 documents and required 87 minutes. The relatively small increase in processing time between 1,099 and 5,731 documents is explained by the fact that overall runtime is dominated not by ConvexTopics clustering itself, but by downstream LLM-based summarization and hierarchical organization steps. These stages depend primarily on the number of generated subtopics and summaries. We are currently working on enabling users to access preliminary results generated by ConvexTopics, allowing them to access the subtopics before the full set of summaries and hierarchical organization is completed.

5 Discussion and Conclusions

BioTopicXplor provides a scalable, reproducible framework for organizing and exploring biomedical literature. It enables exploration not possible with PubMed alone, and is particularly valuable for researchers entering a new area or seeking deeper understanding of a topic. It may also benefit a broader audience by providing a human-interpretable overview of terminology-heavy PubMed literature. We demonstrate its effectiveness through application to the rapidly growing field of anti-aging. Across domains, BioTopicXplor reveals meaningful substructures and maintains consistent topic organization.

Several limitations should be noted. ConvexTopics tends to produce a large number of subtopics, requiring additional post-processing and hierarchical organization to improve interpretability. Additionally, as a web-based system, BioTopicXplor is subject to computational and latency constraints when processing very large document collections in real time. A substantial portion of the processing time is spent generating topic summaries and titles using large language models. These titles are important for comprehension, and play a role in organizing subtopics into higher-level topics. We find that the improved interpretability and resulting organizational structure provided to the user justify the increased processing time.

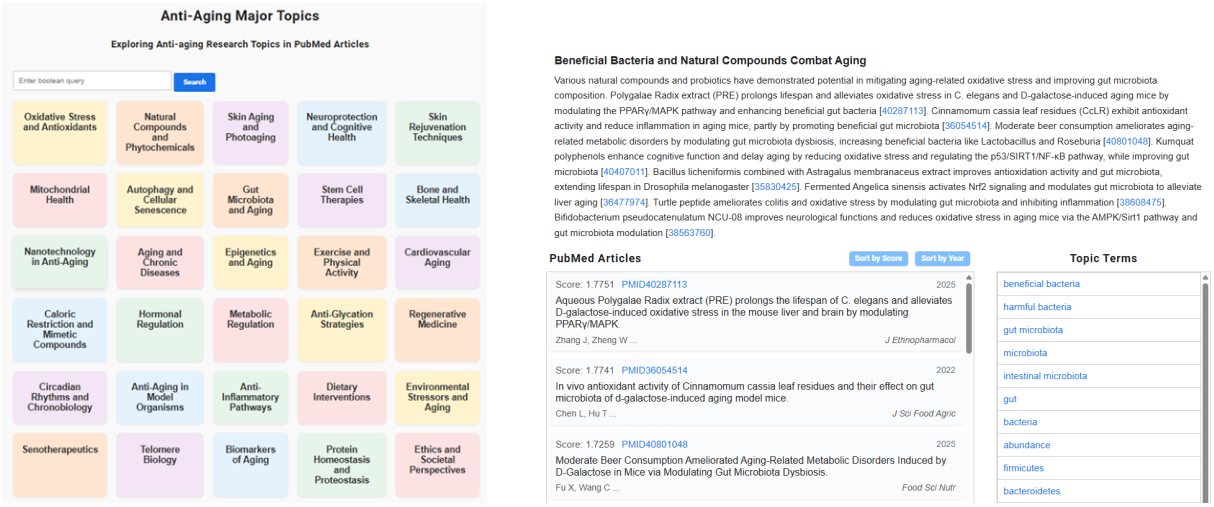

Figure 2: The visualization of the BioTopicXplor applied to the Anti-Aging literature in PubMed. Left panel shows the high-level topics displayed on the main page. Clicking one of these major topics leads to a list of subtopics relevant to the major topic. Clicking on a subtopic leads to a topic page containing topic terms, related pmids, topic summary and title, presented on the right panel. The graphics are adapted from the earlier ConvexTopics paper.

6 Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Andreas Doms and Michael Schroeder. 2005. Gopubmed: exploring pubmed with the gene ontology. *Nucleic acids research*, 33(suppl_2):W783–W786.

Keith D Engwall. 2017. Anne o'tate. *Journal of the Medical Library Association: JMLA*, 105(2):200.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hartigan Ja. 1979. A k-means clustering algorithm. *JR Stat. Soc. Ser. C-Appl. Stat.*, 28:100–108.

Danial Lashkari and Polina Golland. 2007. Convex clustering with exemplar-based models. *Advances in neural information processing systems*, 20.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Harshani Nadeeshani, Jinyao Li, Tianlei Ying, Baohong Zhang, and Jun Lu. 2022. Nicotinamide mononucleotide (nmn) as an anti-aging health product–promises and safety concerns. *Journal of advanced research*, 37:267–278.

Borut Poljšak, Vito Kovač, and Irina Milisav. 2022. Current uncertainties and future challenges regarding nad+ boosting strategies. *Antioxidants*, 11(9):1637.

Qin Song, Xiaofeng Zhou, Kexin Xu, Sishi Liu, Xinqiang Zhu, and Jun Yang. 2023. The safety and antiaging effects of nicotinamide mononucleotide in human clinical trials: an update. *Advances in nutrition*, 14(6):1416–1435.

Theodosios Theodosiou, Konstantinos Vrettos, Ismini Baltsavia, Fotis Baltoumas, Nikolas Papanikolaou, Andreas N Antonakis, Dimitrios Mossialos, Christos A Ouzounis, Vasilis J Promponas, Makrina Karaglani, and 1 others. 2024. Biotextquest v2. 0: an evolved tool for biomedical literature mining and concept discovery. *Computational and Structural Biotechnology Journal*, 23:3247–3253.

Lana Yeganova, Donald C Comeau, Won Kim, and W John Wilbur. 2009. How to interpret pubmed queries and why it matters. *Journal of the American Society for Information Science and Technology*, 60(2):264–274.

Lana Yeganova, Donald C Comeau, and W John Wilbur. 2011. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC bioinformatics*, 12(Suppl 3):S6.

Lana E. Yeganova, Won G. Kim, Shubo Tian, Natalie Xie, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2026. Exploring anti-aging literature via convextopics and large language models. *AMIA Amplify Informatics Conference*.