# Towards Enhanced Topic Discovery on Semantic Maps for Biomedical Literature Exploration

Bin Choi[*]
Yale-NUS College

Brian Ondov[†]
Department of Biomedical
Informatics and Data Science
Yale University

Huan He[‡]
Department of Biomedical
Informatics and Data Science
Yale University

Hua Xu[§]
Department of Biomedical
Informatics and Data Science
Yale University

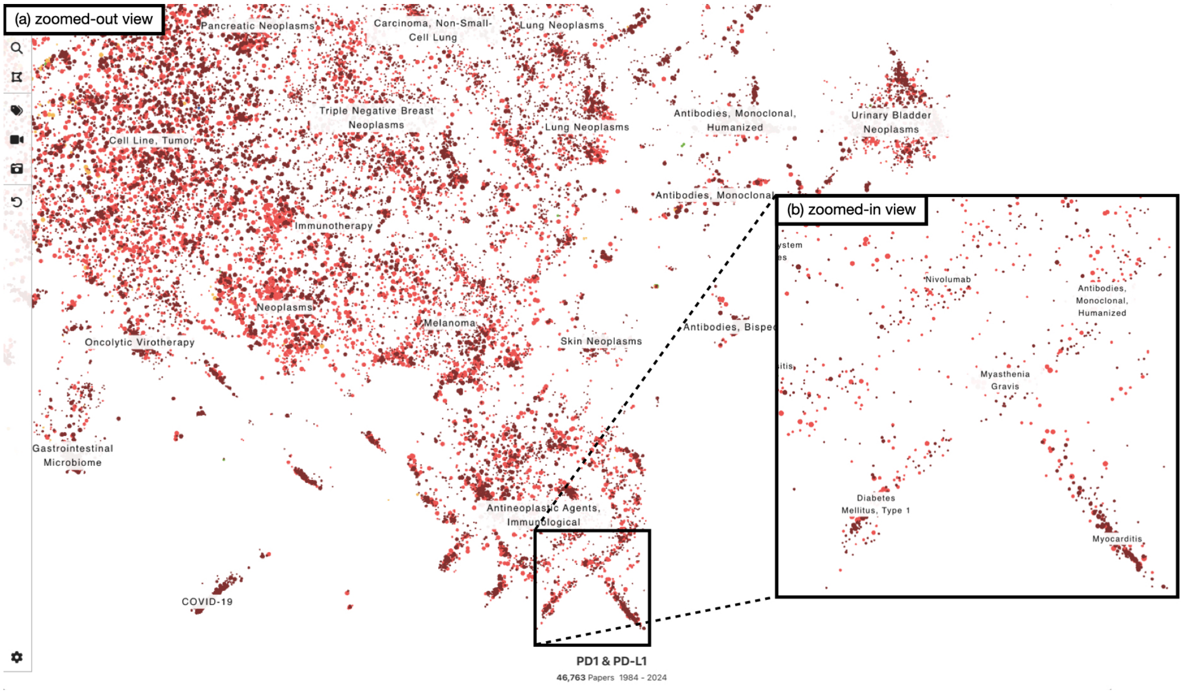

Figure 1: Image of topic labels generated from the semantic map of the PD-1 and PD-L1 literature dataset, displayed on MedViz.org. The figure shows (a) topic labels for high-level clusters and (b) fine-grained subtopic labels within a zoomed-in region.

## Abstract

The rapid growth of biomedical research has led to an overwhelming volume of literature, making it challenging for researchers to efficiently explore and analyze. While existing tools provide an overview of semantic maps and publication distributions, further refinement is needed to reveal fine-grained nuances and hierarchical topics. To address this, we propose a novel method for hierarchical topic modeling and label generation on 2D semantic maps. Our approach consists of three steps. First, we apply density-based hierarchical clustering using HDBSCAN to construct a topic tree. Second, we employ a novel tree-based TF-IDF method to refine topic representation using MeSH terms, capturing both general and local topic distinctions. Finally, we optimize label positioning using a centroid-based method to enhance visualization.

**Index Terms:** Topic modeling, semantic map, tree-based TF-IDF, density-based clustering, data visualization.

## 1 Introduction

The rapid growth of biomedical research has resulted in an exponential increase in the volume of publications, making it essential for researchers to utilize advanced tools to efficiently explore and analyze this vast body of literature [6]. In recent years, literature-mining tools have gained significant attention, helping researchers extract valuable insights and information from large collections of publications. Among these tools, the application of large language models (LLMs) and text embedding techniques has become increasingly pivotal [11]. By converting complex scientific texts into unified high-dimensional vectors, LLMs enable the use of machine learning and data mining techniques to facilitate data analysis and visualization [8].

However, it is still challenging to reveal insights and nuances from large-scale literatures and their high-dimensional vectors, particularly in clustering and topic representation, which are crucial for understanding the topical landscape. In an effective semantic map, closely located embeddings should be semantically related, enabling users to grasp the thematic structure of the map. While existing methods such as BERTopic [7] and Top2Vec [2] offer embedding-based topic modeling, they tend to operate at a fixed level of generality and may not fully capture the fine-grained nuances and hierarchical structures present in large-scale literature collections. To address this, we propose a novel methodology that can capture both broad and specific topics across multiple levels of

[*] e-mail: binchoi@u.yale-nus.edu.sg
[†] e-mail: brian.ondov@yale.edu
[‡] e-mail: huan.he@yale.edu
[§] e-mail: hua.xu@yale.edu

semantic maps for biomedical publications, enabling a more interactive and detailed exploration of topics within clusters and subclusters.

## 2 Methodology

Our proposed method takes a 2D semantic map as input, which is generated from a collection of publications using LLM embedding and dimensionality reduction techniques. In this semantic map, each publication is represented as a point based on its 2D embeddings. We generated the embeddings using transformer-based language model BAAI/bge-small-en-v1.5 [3] and explored various dimensionality reduction techniques — such as LargeVis [12], Uniform Manifold Approximation and Projection (UMAP) [10], and t-distributed Stochastic Neighbor Embedding (tSNE) [13]— to create semantic maps that maintain meaningful spatial relationships. Although our method was tested on maps produced using these specific techniques, other text embedding models and dimensionality reduction approaches can be used, as long as they result in a 2D semantic map that preserves meaningful spatial relationships.

Then, our method conducts three phases to visualize the topic labels for this semantic map: (1) density-based hierarchical clustering, (2) topic representation based on MeSH terms, and (3) label positioning and data file generation. The details of each phase are as follows.

### 2.1 Phase 1: density-based hierarchical clustering

In this phase, we identify groups of semantically similar papers by analyzing their proximity in the semantic space. This process is conducted in a hierarchical manner, beginning with fine-grained clusters and expanding to identify more general clusters. To capture the hierarchical structure of the semantic map, we propose using a topic tree, which organizes and represent topics as a tree structure within the semantic space. Each node in the topic tree represents a cluster of publications. Child nodes represent sub-clusters, reflecting subtopics within the broader context of the parent cluster. In this bottom-up process, we leverage a density-based clustering algorithm HDBSCAN [9], which excels at handling outliers and does not require pre-specifying the number of clusters, to group relevant documents.

Initially, we use heuristically defined parameters to identify fine-grained clusters (i.e., `minimum_cluster_size`, `min_samples`). Then, the algorithm's parameters are incrementally adjusted to identify larger, coarser-grained clusters. As each new, higher-level of clusters are identified, we assign pre-identified sub-clusters to new 'parent' clusters based on membership conditions, such as an overlap threshold. Some fine-grained clusters may not map to a parent cluster and remain as root nodes, representing the highest-level view of their topic. This process continues until an appropriate clustering for the highest-level view of the semantic space is achieved.

### 2.2 Phase 2: topic representation based on MeSH terms, c-TF-IDF, and tree-based TF-IDF

As MeSH terms offer comprehensive coverage of biomedical topics and are organized hierarchically, we use them for topic representation of clusters. Through the first phase, we have identified a hierarchical structure of clusters, where each cluster represents a distinct topic at varying levels of granularity within the semantic map. At the highest level, we can determine the major topics present on the given map using a technique which BERTopic coins as class-based TF-IDF (c-TF-IDF) [7]. In our adaptation, we compile the MeSH terms of each document within a cluster and compare them with those of other clusters to identify the most relevant and representative MeSH term for each cluster.

While this approach works well for representing topics at the root nodes of the topic forest, it struggles to distinguish local differences among sibling sub-clusters further down the topic tree, often resulting in non-discriminative topics. To address this, we propose a novel tree-based TF-IDF procedure that applies c-TF-IDF selectively within sibling (or relative) sub-clusters. Unlike traditional TF-IDF and c-TF-IDF, which calculate Inverse Document Frequency (IDF) across the entire document set, tree-based TF-IDF computes IDF locally within sub-clusters that share a common parent node. This method more effectively identifies local differences and nuances among closely related sibling sub-clusters, as confirmed by our experiments.

### 2.3 Phase 3: label positioning and data file generation

In the final phase, we focus on positioning the generated topic labels on the 2D semantic map and producing structured data outputs to support further analysis and visualization. After determining the representative topic for each cluster in the second phase, we apply a centroid-based approach to place each label at the geometric center of its corresponding cluster. To reflect the hierarchical structure, we assign zoom levels to each label based on its position in the topic tree: higher-level topics remain visible when zoomed out, while more specific subtopics appear only as users zoom in.

Once label positions are decided, we generate a JSON format data file for rendering the labeled semantic map. The file includes the cluster membership of each document, the hierarchical structure of clusters, and the final positions of the topic labels.

## 3 Evaluation

As our research is still in early stages, we have conducted a preliminary evaluation to assess the feasibility of the proposed algorithm. A more comprehensive evaluation is planned for future work.

To validate our method, we implemented the proposed algorithms in Python and developed a prototype based on our existing visualization system using three.js [4] and WebGL techniques [1]. The source code is available on GitHub [5]. Moreover, we created a sample dataset of 46,763 papers related to programmed cell death 1 (PD-1) and programmed death-ligand 1 (PD-L1) proteins from PubMed using the query `PD-1 & PD-L1`. We extracted key metadata from each paper including the title, abstract, MeSH terms, and other relevant information required for our algorithm.

Our method identified three hierarchical levels of topics within the sample dataset. At the top level, 59 broad topics were identified, represented by labels such as *immunotherapy* and *gastrointestinal microbiome*. The second level revealed 200 sub-topics, offering more granular thematic distinctions. At the lowest level, 334 fine-grained topics were identified, with labels like *Myasthenia Gravis*. These topics were rendered on the semantic map, as shown in Fig. 1, where fine-grained clusters emerge as users zoom in on the map (Fig. 1b).

## 4 Discussion and future work

During the development of our method, we demonstrated our prototype to domain experts of immunotherapy and got positive feedback. They commented that the labels are intuitive, but some labels are visually overlapping. In addition, they observed that some smaller clusters, despite representing distinct topics, were still assigned the same labels.

As the next step, we plan to optimize our positioning algorithm by implementing a density-aware approach that adjusts label placement based on the local density and proximity of neighboring clusters. We will also gather more user feedback to thoroughly evaluate our clustering and labeling performance across key metrics such as coherence, accuracy, and robustness. Finally, we will enhance our topic summarization and representation by leveraging large language models to generate topic labels that extend beyond MeSH terms and capture subtle distinctions between clusters.

## Acknowledgments

The authors thank Dr. Xueqing Peng, Dr. Tianxiang Zhang, Dr. Lieping Chen, and other domain experts for providing sample datasets and offering valuable insights during the evaluation of the results throughout the development process.

This work was supported in part by a fellowship grant from Yale College and Yale-NUS College.

## References

[1] E. Angel and D. Shreiner. *Interactive computer graphics with WebGL*. Addison-Wesley Professional, 2014. 2

[2] D. Angelov. Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*, Aug 2020. doi: 10.48550/arXiv.2008.09470 1

[3] BAAI. bge-small-en-v1.5. https://huggingface.co/BAAI/bge-small-en-v1.5, 2023. Hugging Face. 2

[4] R. Cabello. Three.js. https://github.com/mrdoob/three.js, 2010. 2

[5] B. Choi. Vahc topic modeling. https://github.com/binchoi/vahc-topic-modeling, 2024. 2

[6] R. González-Márquez, L. Schmidt, B. M. Schmidt, P. Berens, and D. Kobak. The landscape of biomedical research. *Patterns (N Y)*, 5(6):100968, Apr 2024. doi: 10.1016/j.patter.2024.100968 1

[7] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, Mar 2022. doi: 10.48550/arXiv.2203.05794 1, 2

[8] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. Insightpilot: An llm-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 346–352, 2023. 1

[9] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205 2

[10] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2

[11] E. Simon, K. Swanson, and J. Zou. Language models for biological research: a primer. *Nat Methods*, 21:1422–1429, 2024. doi: 10.1038/s41592-024-02354-y 1

[12] J. Tang, J. Liu, M. Zhang, and Q. Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297. International World Wide Web Conferences Steering Committee, 2016. 2

[13] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2