

# Interactive Distillation of Large Single-Topic Corpora of Scientific Papers

Nick Solovyev

*Theoretical Division, LANL*

Los Alamos, USA

nks@lanl.gov

Ryan Barron

*Theoretical Division, LANL*

Los Alamos, USA

barron@lanl.gov

Manish Bhattarai

*Theoretical Division, LANL*

Los Alamos, USA

ceodsppectrum@lanl.gov

Maksim E. Eren

*Advanced Research in Cyber Systems, LANL*

Los Alamos, USA

maksim@lanl.gov

Kim Ø. Rasmussen

*Theoretical Division, LANL*

Los Alamos, USA

kor@lanl.gov

Boian S. Alexandrov

*Theoretical Division, LANL*

Los Alamos, USA

boian@lanl.gov

**Abstract**—Highly specific datasets of scientific literature are important for both research and education. However, it is difficult to build such datasets at scale. A common approach is to build these datasets reductively by applying topic modeling on an established corpus and selecting specific topics. A more robust but time-consuming approach is to build the dataset constructively in which a subject matter expert (SME) handpicks documents. This method does not scale and is prone to error as the dataset grows. Here we showcase a new tool, based on machine learning, for constructively generating targeted datasets of scientific literature. Given a small initial “core” corpus of papers, we build a citation network of documents. At each step of the citation network, we generate text embeddings using the transformer generated science-specific large language model SciNCL [Ostendorff, Malte, et al. “Neighborhood contrastive learning for scientific document representations with citation embeddings.” arXiv preprint arXiv:2202.06671 (2022).] and visualize the embeddings through dimensionality reduction. Papers are kept in the dataset if they are “similar” to the core or are otherwise novelly pruned through human-in-the-loop selection. Additional insight into the papers is gained through sub-topic modeling using SeNMFk. We demonstrate our new tool for literature review by applying it to two different fields in machine learning.

**Index Terms**—transformers, nlp, non-negative matrix factorization, data visualization

## I. INTRODUCTION

One of the integral tasks of scientific research is the literature review of highly specific topics of interest. Literature review often involves identifying papers of interest based on keyword searches and following the relevant citations. This manual process, however, is prone to miss potentially significant papers and information. In addition, organizing highly specific scientific literature datasets and applying data analysis techniques, such as topic modeling, may allow a deeper understanding of a given field and the discovery of new research directions. However, curating such highly-specific datasets of scientific literature requires the time-consuming help of a subject matter expert (SME). Here, we introduce a new assistant tool based on machine learning (ML) that allows for building highly-specific scientific literature datasets. Bibliographic Utility Network Information Expansion (BUNIE)

streamlines the literature review task with a user-friendly and intuitive system while enhancing the specificity of the papers of interest using ML techniques and integrated human-in-the-loop procedures.

In this work, we contribute a novel approach to the scientific dataset expansion problem by jointly integrating Transformer-based document text embeddings with human-in-the-loop pruning to generate targeted scientific datasets. We then use non-negative matrix factorization (NMF) with automatic model determination (NMFk) for modeling the topics in these papers to further refine our datasets [1]. Our approach is unique in its inclusion of a human-in-the-loop for enhancing and distilling the extracted topics, such that the corpus of papers is narrowed down via an interactive process. To the best of our knowledge, this iterative method is the first of its kind to offer users the ability to analyze the topic modeling results and apply their feedback to enhance the literature review procedure by steering the ML output. The feedback loop enables the users to grow and refine the results until a targeted dataset of a specific size is reached, providing a unique and interactive solution to large-scale literature review.

The process begins with a small number of core papers selected from a topic of interest by an SME. At this initial stage, the topic may not fully align with the user’s specific objectives and is likely incomplete. The core papers are used as a reference to obtain an additional set of relevant documents that increase the size and enhance the specificity of the existing dataset. The additional documents are selected using a citation network formed from the existing papers in the dataset. The expansion results are then pruned using multiple methods, including an interactive selection by the user, document embedding similarity metrics, and topic modeling. In contrast to the traditional static approach of computing the topics, our approach is iterative and dynamic. It allows repetition of this refinement cycle, growing the dataset with each iteration. This enables the creation of large but specific datasets, ideal for training large language models. Through this interactive, user-driven approach, we empower users to steer the topic

extraction process directly, ensuring the results are tailored to their specific requirements. This paper demonstrates our novel tool by exploring the scientific literature on applying tensor decomposition for numerical solutions of partial differential equations.

Our contributions include:

- Introducing a novel paper selection and visualization tool for scientific dataset curation and literature review.
- Utilizing text embeddings together with dimensionality reduction techniques to model the documents.
- Integrating our machine learning approach to scientific literature with human-in-the-loop procedures for refining and guiding text modeling.
- Demonstrating the capabilities of our tool by applying it to the scientific literature in two different scientific fields.

## II. RELATED WORK

This section summarizes techniques and prior works applied to forging a highly-specific dataset of research papers.

1) *Topic Modeling & Tensor Decomposition*: A common approach to topic modeling is through Non-Negative Matrix Factorization (NMF) [2], [3]. NMF can be applied to a words by documents matrix to identify latent patterns within the corpus. An extension to NMF, Semantic NMF with automatic model determination (SeNMFk), is leveraged in [4]–[6] to perform topic modeling while incorporating the text’s semantic structure. The aforementioned documents by word matrix used in NMF and SeNMFk is usually the term frequency–inverse document frequency (TF–IDF) matrix together with the co-occurrence/word-context matrix, the values of which represent the number of times two words co-occur in a predetermined window of the text. This is a common method of vectorization, however, more advanced methods to process the documents exist.

2) *Document Embeddings & Transformers*: Vector representations of a text were previously used for dimensional mapping, cross-comparisons, and similarity analysis [7]. Common models for learning word embeddings have been Global Vectors for Word Representation (GloVe) [8] and Word2Vec [9]. Recently, transformers have been used for large language models (LLM) as internal states, as well as for topic modeling [10]–[13]. A popular transformer-based LLM is the Bidirectional Encoder Representations from Transformers (BERT) [14]. An example of BERTs topic modeling and sentiment analysis is discussed in Ref. [15], where emotions related to the use of ChatGPT [16] are studied through social media posts by medical field researchers. While transformers, as used in BERT are useful, they have limitations, as demonstrated in Ref. [17], which addresses BERT’s few hundred-word input capacity by adding transformer layers to segment text, paired with two activation layers for final classification. In our work, we apply the SciNCL transformer to generate text embeddings [18]. Citation data is used as an additional training signal in SciNCL’s document embedding closeness, aiding the document distance determinations.

3) *Data Visualization and Tools*: A range of tools is available to explore and analyze research papers. For instance, citation network and topic modeling tools such as Topic Modeling Tool [19], and Stanford Topic Modeling Toolbox [20] are publicly available. While these tools excel in gathering topical data from their inputs, they lack visual representation. Another tool with visualizations, ‘Connected Papers,’ serves as a resource for discovering scientific literature [21]. From single document inputs, Connected Papers produces a graph where each paper is a node, positioned according to a coupling of the co-citations and bibliography rather than direct citations [21]. While ‘Connected Papers’ is useful for exploring scientific literature based on the bibliography information, our tool advances the utility of bibliography information by creating a specialized dataset of documents leveraging a citation network coupled with human-in-the-loop and machine learning procedures. Another research paper visualization tool, designed specifically for the influx of Covid-19 papers at the height of the pandemic, is explained in Ref. [22]. Within this tool, the process begins with text cleaning (tokenization, removal of stop-words, & punctuation & capitalization), transformation into a TF-IDF matrix, then t-distributed stochastic neighbor embedding (t-SNE) [23] reduced dimensions for graphing. Here, we use Uniform Manifold Approximation and Projection (UMAP) [24] to reduce the 768-dimensional embeddings output by SciNCL [18] to a two-dimensional projection.

4) *Human in the Loop*: Incorporating user feedback into the systems has seen recent adoption into several schemes, including OpenAI’s ChatGPT [16] and Google’s BARD [25]. In the study Ref. [26], a knowledge graph is built by the framework textually prompting a user, collecting feedback in every response to provide an acceptable retail-item recommendation. Significant differences between BUNIE and Ref. [26] exist. For instance, the study’s structure provides one recommendation, whereas BUNIE offers an entire dataset. Interactive modes also differ. Our system uses click-and-drag selection to delete papers rather than a textual conversation. Furthermore, BUNIE only removes papers at the HITL phase until further citation hops are requested for more documents. Contrastly, the tool described in Ref. [26] requests positive and negative feedback about recommendations. A HITL work more similar to BUNIE is described in Ref. [27], which aims to build labeled image datasets for Computer Vision (CV) applications. A user labels a few images, which extrapolate to all in the image’s cluster, then the images are evaluated by a model for reassignment. Like BUNIE, the process is iterated to convergence but differs in direct user influence of datum retention. Moreover, HITL has been examined from the perspective of assisting artificial intelligence (AI) in Ref. [28], specifically for natural language processing (NLP), CV, an NLP and CV pairing, and real-world robotic applications. Still, HITL from the perspective of non-robotic real-world applications supported by NLP is not considered. In our tool, a user inputs one or more documents, iteratively grows the data, and deletes documents at every iteration, hand-in-hand with both AI and tensor pruning.

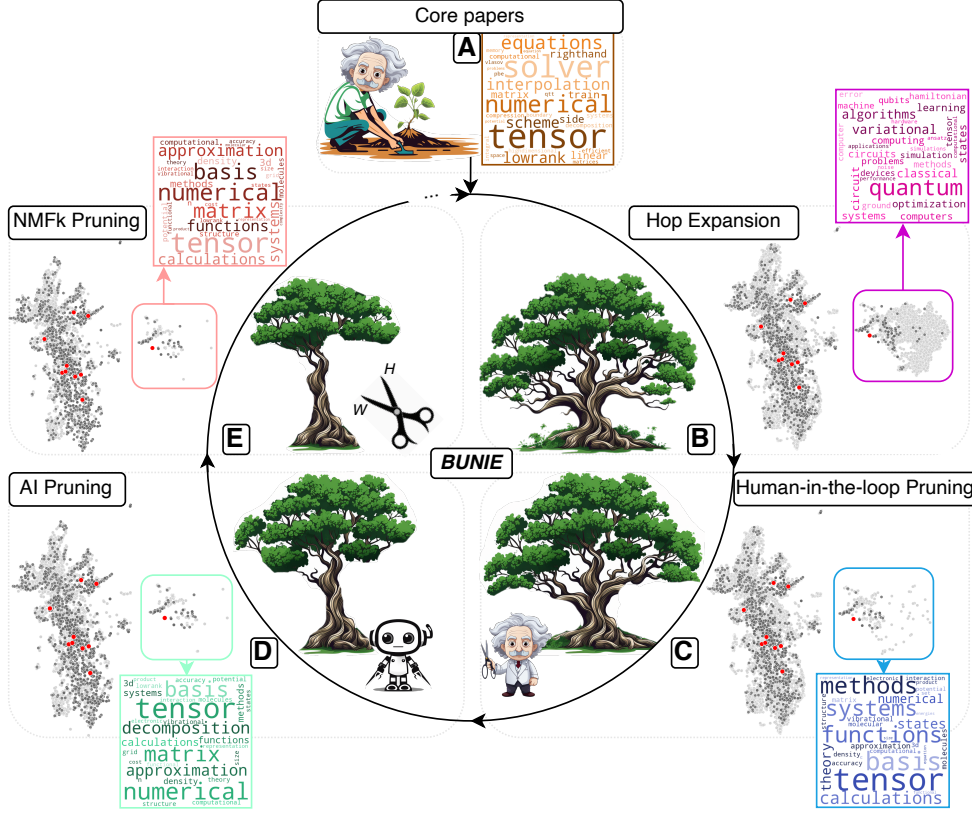


Fig. 1: Illustration of BUNIE’s distillation pipeline. **Panel A** inputs to BUNIE the SME-selected, highly-specific papers (core), where a subset is represented as a bag-of-words wordcloud. **Panel B** dataset is expanded through the core’s citation network. A subset of papers is selected to showcase how much the content differs from the core **Panel C** features human-in-the-loop pruning. Reduction in the subset cluster is visible and the wordcloud begins to resemble the core. **Panel D** documents are pruned through a document embedding heuristic, removing papers too far from the core in the embedding space. The dataset becomes more compact and the subset begins to approach wordcloud parity with the original papers. **Panel E** topic modeling through SeNMFk further prunes the dataset. H-clustering the factorization removes clusters that lack core papers, producing to a neatly trimmed tree. This final stage has a dense set yet numerous documents indicating a successful distillation, evidenced in a wordcloud closely resembling the original. Additional cycles can be made for refinement or the data extracted in repose for downstream analysis, the option as the **ellipse**.

### III. METHOD

The utility of BUNIE comes from the combination of being able to quickly expand a dataset of publications by traversing the citation network in combination with being able to curate the dataset at scale by effectively using document text embeddings. As depicted in Figure 1, the workflow is cyclical as it involves iterative steps of acquiring new papers through the citation network and then refining this expanded dataset using various pruning techniques. The ultimate goal is to create an extensive collection of scientific literature centered around a specific topic, using a small, hand-picked set of relevant papers as the starting point.

1) **Selecting the Core:** First, the user provides BUNIE with a set of “core” papers, comprised of a unifying theme or topic, as the foundation of the dataset. A subject matter

expert (SME) should select and/or review the core for the best results to ensure quality and relevance. It is important to remember that BUNIE expands the dataset by traversing the citation network of the core papers. A single, well-cited document may produce an extensive dataset after a few iterations/hops following the citation network, while a collection of less frequently cited documents might yield a more limited network. In our experiments, BUNIE has been successfully applied to cores ranging from as few as 6 documents and as many as 63 documents. The core papers are inputted into BUNIE using unique paper identifiers such as DOI. Using the SemanticScholar API [29], BUNIE extracts basic information about these documents, including the title, abstract, year of publication, authors, citations, and references.

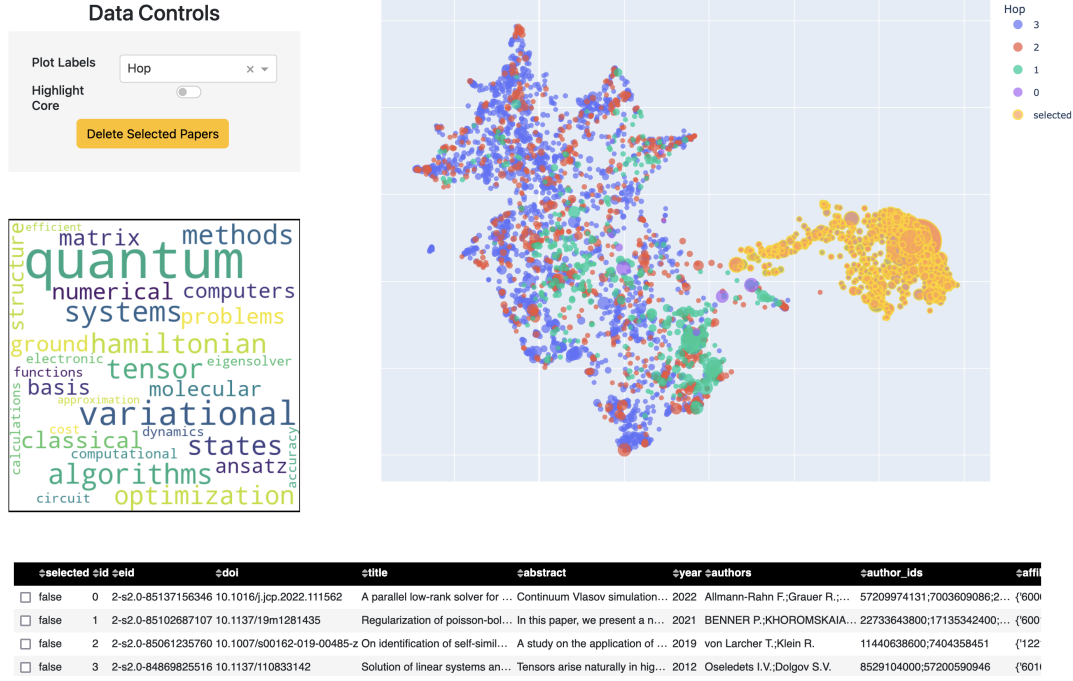


Fig. 2: Screenshot of the GUI built for BUNIE. The user is able to upload core papers, perform hops, and do the HITL pruning. This screenshot shows a user pruning a 3-hop dataset. The papers on the right-hand side of the plot (highlighted in yellow) have been manually selected, and the bag-of-words wordcloud for the selection is shown to the left. The *Hop* legend represents the number of hops (hop 0 is the core set of documents), and the SME selected/highlighted documents.

2) **Expanding the Dataset:** With the core established, the user can grow the dataset by making a "hop" within the citation network. The citation network represents a directed graph formed by publications and their respective citations. If we denote a document,  $a$ , as belonging to a set of documents  $X$  and a document,  $b$ , belonging to the set of their citation  $X^c$ , we can say that  $a \rightarrow b$  if and only if  $b$  cites  $a$ . In this context, a hop can be defined as  $X := X \cup X^c$ . In this fashion, a second hop would also incorporate the citations from the documents in  $X^c$ , which was acquired from the first hop. The number of hops performed is left to the user's discretion, thereby controlling the scale of dataset expansion. The process can continue until the dataset reaches a desired size or until the entire citation network has been traversed. BUNIE also offers the capability to form the citation network with the edges reversed, using references as the basis instead of citations. This feature can be particularly useful when the core consists of relatively new or infrequently cited publications.

3) **Pruning the Dataset:** Given the interconnected nature of the citation network, not every paper found through the hop process will be relevant to the core. For example, a highly influential publication may be cited as an acknowledgment in subsequent studies focusing on entirely new issues. Thus, it is crucial to perform pruning at each hop along the citation network to prevent irrelevant topics from propagating within the growing dataset. In BUNIE, pruning is accomplished through a combination of the following three techniques.

#### A. Human in the Loop (HITL) Pruning:

Textual similarity comparison presents a substantial challenge for humans and computational algorithms. To simplify this task, we employ SciNCL [18] to transform the aggregated titles and abstracts of the dataset into 768-dimensional embeddings. These high-dimensional embeddings are reduced to a two-dimensional projection using UMAP [24]. Semantically similar papers tend to cluster together in this two-dimensional space when plotted on a scatter plot, providing an intuitive visual representation of the dataset's structure. Although not perfect, this process simplifies manual content comparison.

To aid in the manual analysis and document pruning, we have designed a graphical user interface (GUI) to quickly select and examine many papers using the UMAP visualized projection of the embeddings, as shown in Figure 2. The SME can highlight papers by drawing a custom lasso or rectangle over the projected papers. The tool then generates a bag-of-words wordcloud to show the most frequent vocabulary in the chosen paper set. For a finer-grain analysis, the GUI provides a data table displaying all known data fields for the selected papers that an SME can analyze.

#### B. Automatic Pruning of Document Embeddings

As with any dimensionality reduction algorithm, UMAP necessitates tradeoffs in the data's representation in the two-dimensional space. While useful for document visualization and enabling HITL pruning, a significant portion of the



embedding structure is lost. To counteract this loss, we introduce a method for pruning the document embeddings in their original high-dimensional space. Each core paper in the dataset is considered specialized within its field. Therefore, the embeddings of the new papers added through the citation network are evaluated for their proximity to each of the core paper embeddings. The intuition is that the embeddings of relevant papers should reside "close" to one or more of the embeddings of the core papers. Each core embedding is treated as the center of a hypersphere with radius  $\rho$ .

In mathematical notation, given a set of core papers  $C = \{c_1, c_2, \dots, c_n\}$  and their corresponding embeddings  $E = \{e_1, e_2, \dots, e_n\}$ , the radius  $r$  for each hypersphere is calculated as: *First*, compute the pairwise Euclidean distances for all embeddings in  $E$ , forming a set  $D = d(e_i, e_j) : e_i, e_j \in E, i \neq j$  where  $d(e_i, e_j)$  denotes the Euclidean distance between embeddings  $e_i$  and  $e_j$ . *Second*, the median Euclidean distance between all core embeddings, denoted as  $\rho$ , where  $\rho = \text{median}(D)$ , can then be used as a threshold for including newly cited documents. The embedding for each core paper becomes a center of a hypersphere with radius  $\rho$ . A document embedding within one or more hyperspheres is at least as close to one or more core document embeddings as the median separation of the core document embeddings. Figure 3 illustrates the process simplified to a 2-dimensional space.

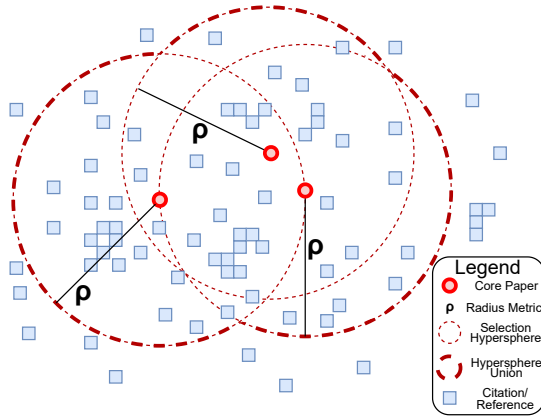


Fig. 3: Hypersphere pruning calculation. Radii are compared from core paper distances, and once  $\rho$  is selected, papers beyond the perimeter are pruned.

### C. Pruning through Topic Modeling

To further ensure topic cohesion, we perform topic modeling on the pruned dataset formed in the previous two steps. We utilize Semantic non-negative matrix factorization with automatic model selection (SeNMFk) [30] for topic modeling. Given the documents dataset, we form a term frequency-inverse document frequency (TF-IDF) matrix  $\mathbf{X} \in \mathbb{R}_+^{m \times n}$  and an SPPMI matrix  $\mathbf{S} \in \mathbb{R}_+^{m \times m}$  which encodes the semantic structure of the data (where  $m$  is the number of tokens in the vocabulary and  $n$  is the number of documents). We then jointly factorize  $\mathbf{X}$  and  $\mathbf{S}$  to produce two non-negative factor matrices

$\mathbf{W} \in \mathbb{R}_+^{m \times k}$  and  $\mathbf{H} \in \mathbb{R}_+^{k \times n}$ , such that  $\mathbf{X}_{ij} \approx \sum_s^k \mathbf{W}_{is} \mathbf{H}_{sj}$ . Here,  $\mathbf{W}$  represents the distribution of words across different topics, and  $\mathbf{H}$  describes how these topics are distributed across the documents. After applying NMF, the information in  $\mathbf{H}$  is used to associate each document with the topic it contributes most, forming clusters of documents. Only the documents corresponding to the topic, which comprises the core documents, are preserved.

It is important to conduct robust pre-processing of the documents to limit the noise introduced by expanding the dataset to produce a meaningful vocabulary. Our pre-processing procedure removes common stop-words, symbols, newline characters, HTML tags, non-ASCII characters, e-mail addresses, and copyright statements. Documents in languages other than English are identified and removed using heuristics such as the ratio of non-ASCII to total characters and the occurrence of common English stop-words in the text. There are instances where specific tokens or phrases denote unique terms in the chosen domain. While these terms might appear in different forms (such as spelling, acronym, or hyphenation), all forms signify the same concept. Standard preprocessing may split a multi-token term into separate tokens, which can destroy potentially crucial meaning. However, given that an SME initially chooses the core papers, the SME can also pinpoint important terms and their assorted forms. Once these terms are identified, we consolidate all forms of each term into a singular entity. In the case of multi-token terms, we retain either the acronym or a hyphenated version to ensure that the term's meaning is preserved in the TF-IDF and SPPMI matrices. In our tensors literature example, we substitute *tensor-train* with  $\{TT, \text{tensor train}\}$  and *partial-differential-equation* with *PDE* and all other various forms. Another strategy we employ at this pruning step to reduce noise involves reusing the same vocabulary for every hop. The vocabulary, derived from the core papers is consistently applied at each pruning decomposition. Consequently, less relevant papers (those using a significantly different vocabulary than the core) are represented as sparse entries in the TF-IDF matrix, reducing their influence on the decomposition. This step also enhances computational efficiency as the vocabulary dimension remains constant and does not grow with the number of documents.

Through these methods, BUNIE effectively enhances the thematic coherence of the dataset while maintaining topical alignment with the original core. This results in a significantly larger, interconnected dataset that retains the integrity of the original subject matter, ready for more in-depth exploration or application. Furthermore, to quantify the efficacy of our approach, we employed a compactness score, which is a metric that evaluates how closely the documents in the dataset are related to each other in terms of the topics they cover. The compactness score of a dataset is calculated using cosine similarity between the document embeddings. In mathematical terms, given a set of document embeddings  $E = \{e_1, e_2, \dots, e_n\}$ , the compactness score  $C$  is given by:

$$C = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \frac{e_i \cdot e_j}{|e_i|_2 |e_j|_2} \quad (1)$$

where  $n$  is the total number of documents,  $e_i$  and  $e_j$  are the embeddings of the  $i^{th}$  and  $j^{th}$  document,  $\cdot$  denotes the dot product, and  $|\cdot|_2$  denotes the Euclidean norm. In measuring topic coherence using document embeddings, the cosine similarity between two embeddings, which ranges between -1 and 1, provides a measure of semantic alignment. A negative cosine similarity score, implying that the documents are semantically opposed, is an unlikely scenario within a specific topic. Therefore, we constrain the compactness score to fall between 0 and 1 to facilitate a meaningful quantification of topic coherence or alignment, accomplished by taking the absolute value of the cosine similarity. Higher values suggest a greater topic similarity between documents. The final compactness score, a value also ranging between 0 and 1, is computed as the average cosine similarity across all pairs of documents in the dataset. By this measure, a higher compactness score indicates a more coherent or well-aligned set of documents regarding their topical content.

#### IV. RESULTS

This section presents two experimental uses of BUNIE.

##### A. Expanding Targeted Dataset

We first applied BUNIE to 10 papers hand-picked by an SME on a specific topic. These publications were influential papers in solving integral equations using tensor-train decomposition. With the "core" established, we sought to expand the dataset along the citation network. After the first hop, 632 citing papers were found. Using the visualization tool, we could quickly locate and prune papers that did not match the topic. While these papers tangentially addressed tensor decomposition, they failed to engage with the specific issues highlighted in the core papers. We then applied the automatic pruning through embeddings and SeNMFk pruning. After pruning the first hop, we were left with 411 papers, including the original 10 core papers.

For such a minimal subset of papers, it was feasible to use the two-dimensional projection of the document embeddings in conjunction with bag-of-words word clouds to promptly identify the outlying papers. However, upon the second citation network expansion, the dataset grew rapidly to more than 8,000 papers. At this stage, the automatic pruning of the citation network became paramount. After pruning the second hops papers, a third hop was performed. After pruning, the final result came to 3,915 papers. This data flow demonstrates how BUNIE effectively combines human intuition with algorithmic utility to create a focused, relevant scientific dataset.

As demonstrated in Table I, BUNIE's iterative process of topic expansion and alignment increases the compactness score of the dataset. While the first expansion to the citation network added many new documents to the dataset, it also introduced many unrelated documents, causing the compactness

score to drop from 0.894 to 0.823. The subsequent automatic pruning based on hypersphere proximity to the core document embeddings was effective in increasing the compactness score to 0.860, by eliminating less relevant documents, reducing the total document count to 4625. Following the hypersphere pruning, we carried out topic alignment by applying SeNMFk decomposition and selecting relevant subtopics, further refining the dataset. This increased the compactness score and resulted in a more manageable dataset containing 3915 documents. The increase in compactness score at each stage of the BUNIE process demonstrates the method's effectiveness in maintaining topic cohesion while expanding the dataset from a small set of core papers.

TABLE I: Compactness Score - Tensors

Dataset	Compactness	Num. Documents
Core Papers	0.894	10
3-Hops, No Pruning	0.823	10338
3-Hops, After Hypersphere Pruning	0.860	4625
3-Hops, After SeNMFk Pruning	0.861	3915

TABLE II: Compactness Score - Audio processing

Dataset	Compactness	Num. Documents
Core Papers	0.913	64
4-Hops, No Pruning	0.798	15294
4-Hops, After Hypersphere Pruning	0.861	1987
4-Hops, After SeNMFk Pruning	0.861	1081

##### B. Exploratory Data Expansion

In recent years, the paper "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context" [31] has drawn significant attention and influence across multiple research domains. Given this impact, it becomes interesting to explore the different domains influenced by the paper either individually or in relation to each other. BUNIE allows us to perform this exploration through topic modeling and visualizing text embedding projections.

As demonstrated in 4b, we identified five prominent clusters associated with the following topics: audio processing, computer vision, speech processing, natural language processing, and proteins. From the visualized clusters, cluster I, containing music terms from 68 papers, was expanded through four hops along the citation and reference network, resulting in a significantly larger dataset comprising 15,294 papers. Following the expansion, the dataset was pruned through hypersphere calculation, retaining only papers within at least one of the 64 first-hop paper hyperspheres. At this point, the dataset contained 1,987 papers. Next, SeNMFk decomposed the papers into their core topic clusters, preserving 1,081 papers through 19 clusters, where only eight contained core papers and were preserved as the final dataset. The top words from the retained clusters in order were: music, attention, generative, lyric, video, score, learn, and emotion. As Table II shows, the compactness of the dataset increased with each

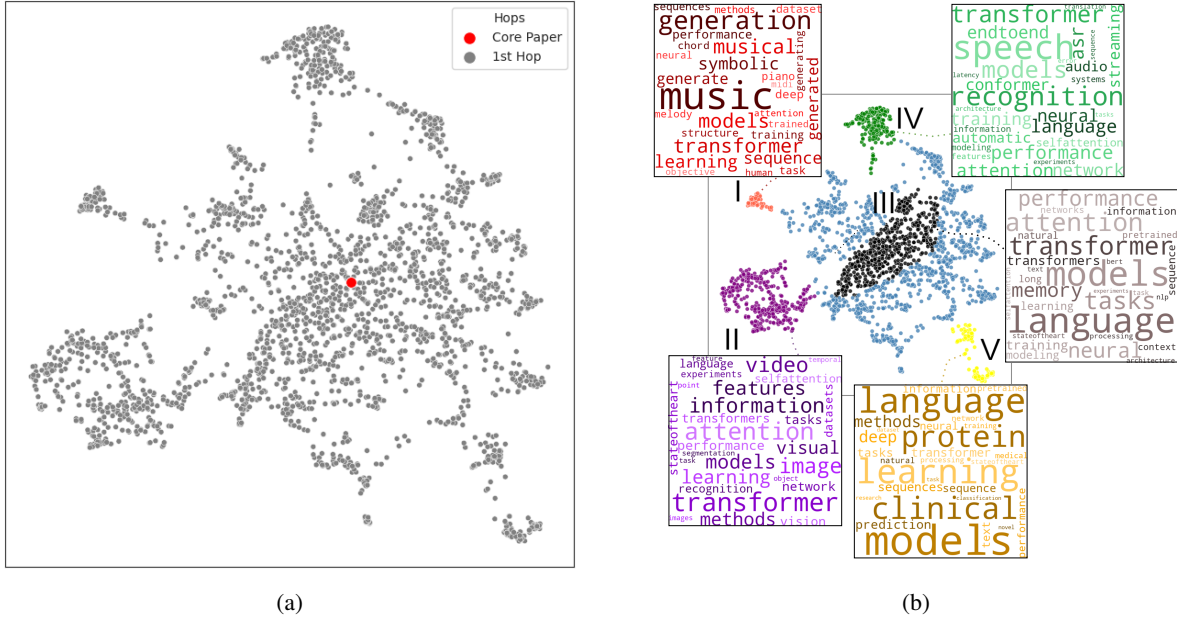


Fig. 4: Exploration of reference and citation paper topics out of an influential transformer paper  
(a) “Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context” in red, citations/references in grey  
(b) Manually selected paper clusters & wordclouds from 4a colored: Music, Video, Transformer Performance, Speech, Proteins

pruning step. The compactness of the core papers was 0.913, which decreased to 0.798 after the four-hop expansion due to the introduction of less-relevant papers. After hypersphere pruning, the compactness increased to 0.861, indicating the successful removal of off-topic papers. Remarkably, the compactness remained stable after SeNMFk pruning, suggesting that the most relevant papers were retained.

Notably, retained paper distributions per hypersphere pruned embedding mappings and SeNMFk decompositions will not always align with a human curator’s intuitive UMAP-reduced selections. The discrepancy highlights the unique value of human judgment with algorithmic tools in dataset curations.

## V. CONCLUSION

This work contributes a novel system to build scientific datasets. With minimal input, we are able to iteratively build a dataset of scientific literature anchored on the core subject provided by an SME. At each step, the dataset is enlarged through the citation network and subsequently pruned using three separate methods, including one with human-in-the-loop. The result is an expanded dataset of work relevant to the core.

Promising future work is to seed an initial topic specification. The system would then iterate autonomously, filtering out documents and recalculating topic estimates to achieve topic distillation based on reinforcement learning. Auto-distillation could dynamically adapt the topic extraction and refinement based on continuous feedback on the topic’s state. The system’s efficiency and accuracy could improve over time, leading to more precise and reliable topic distillation.

Additional considerations for future work include methods of forming the embeddings and creating a ‘synthetic’ core

paper to serve as a foundation for automated topic alignment. The utilization of graph neural networks for understanding the relationship between the citations can also be explored, offering further insights into the structure and interconnections of the scientific literature. These enhancements and additions would augment the effectiveness and flexibility of BUNIE, further assisting researchers in their quest for knowledge.

## VI. ACKNOWLEDGMENT

This research was funded by DOE National Nuclear Security Administration (NNSA) - Office of Defense Nuclear Nonproliferation R&D NA-22 grant DE-AC52-06NA2539 supported by Los Alamos National Laboratory’s Institutional Computing Program, and by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

## REFERENCES

- [1] B. S. Alexandrov, V. Vesselinov, and K. Ø. Rasmussen, “SmartTensors unsupervised AI platform for big-data analytics,” Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2021, LA-UR-21-25064. [Online]. Available: <https://www.lanl.gov/collaboration/smart-tensors/>
- [2] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, pp. 111–126, 1994.
- [3] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” 07 2003, pp. 267–273.
- [4] R. Vangara, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, M. Bhattarai, V. G. Stanev, and B. S. Alexandrov, “Semantic nonnegative matrix factorization with automatic model determination for topic modeling,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 328–335.

- [5] R. Vangara, M. Bhattarai, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, V. G. Stanev, and B. S. Alexandrov, "Finding the number of latent topics with semantic non-negative matrix factorization," *IEEE Access*, vol. 9, pp. 117 217–117 231, 2021.
- [6] M. E. Eren, N. Solovyev, M. Bhattarai, K. Ø. Rasmussen, C. Nicholas, and B. S. Alexandrov, "Senmfk-split: Large corpora topic modeling by semantic non-negative matrix factorization with automatic model selection," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, ser. DocEng '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: <https://doi.org/10.1145/3558100.3563844>
- [7] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 957–966. [Online]. Available: <https://proceedings.mlr.press/v37/kusnerb15.html>
- [8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.
- [10] K. Li, Z. Liu, T. He, H. Huang, F. Peng, D. Povey, and S. Khudanpur, "An empirical study of transformer-based neural language model adaptation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7934–7938.
- [11] K. Lo, Y. Jin, W. Tan, M. Liu, L. Du, and W. Buntine, "Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence," 2021.
- [12] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," 2021.
- [13] A. Glazkova, *Identifying Topics of Scientific Articles with BERT-Based Approaches and Topic Modeling*, 05 2021, pp. 98–105.
- [14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [15] S. V. Praveen and V. Vajrobol, "Understanding the perceptions of healthcare researchers regarding chatgpt: A study based on bidirectional encoder representation from transformers (bert) sentiment analysis and topic modeling," *Annals of biomedical engineering*, 2023.
- [16] OpenAI, "GPT-3.5-based ChatGPT," 2021. [Online]. Available: <https://openai.com>
- [17] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 838–844.
- [18] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, "Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings," in *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*. Abu Dhabi: Association for Computational Linguistics, December 2022, 7–11 December 2022. Accepted for publication.
- [19] J. S. Enderle, "Topic modeling tool," <https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html>, 2023, accessed: June 1, 2023.
- [20] E. R. Daniel Ramage, "Stanford topic modeling toolbox," <https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4/>, 2009, accessed: June 1, 2023.
- [21] E. Alex, E. Smolyansky, I. Harpaz, and P. Sahar, "Connected papers," <https://www.connectedpapers.com>, 2023, accessed: June 1, 2023.
- [22] M. E. Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "Covid-19 kaggle literature organization," in *Proceedings of the ACM Symposium on Document Engineering 2020*, ser. DocEng '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: <https://doi.org/10.1145/3395027.3419591>
- [23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [24] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
- [25] Google, "Google Bard," 2023. [Online]. Available: <https://bard.google.com/>
- [26] Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. de Melo, and Y. Zhang, "Hoops: Human-in-the-loop graph reasoning for conversational recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2415062421. [Online]. Available: <https://doi.org/10.1145/3404835.3463247>
- [27] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," 2016.
- [28] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, vol. 135, pp. 364–381, oct 2022. [Online]. Available: <https://doi.org/10.1016%2Ffuture.2022.05.014>
- [29] R. M. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, M. Crawford, D. Downey, J. Dunkelberger, O. Etzioni, R. Evans, S. Feldman, J. Gorney, D. W. Graham, F. Hu, R. Huff, D. King, S. Kohlmeier, B. Kuehl, M. Langan, D. Lin, H. Liu, K. Lo, J. Lochner, K. MacMillan, T. Murray, C. Newell, S. Rao, S. Rohatgi, P. L. Sayre, Z. Shen, A. Singh, L. Soldaini, S. Subramanian, A. Tanaka, A. D. Wade, L. M. Wagner, L. L. Wang, C. Wilhelm, C. Wu, J. Yang, A. Zamarron, M. van Zuylen, and D. S. Weld, "The semantic scholar open data platform," *ArXiv*, vol. abs/2301.10140, 2023.
- [30] M. E. Eren, N. Solovyev, M. Bhattarai, K. Ø. Rasmussen, C. Nicholas, and B. S. Alexandrov, "Senmfk-split: large corpora topic modeling by semantic non-negative matrix factorization with automatic model selection," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, 2022, pp. 1–4.
- [31] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 01 2019, pp. 2978–2988.