Demo: Interactive Visualization of Semantic Relationships in a Biomedical Project's Talent Knowledge Graph

Jiawei Xu
School of Information
University of Texas at Austin
jiaweixu@utexas.edu

Zhandos Sembay
School of Medicine
University of Alabama at Birmingham
zsembay8@uab.edu

Swathi Thaker
School of Medicine
University of Alabama at Birmingham
snthaker@uab.edu

Pamela Payne-Foster
School of Medicine
University of Alabama
pfoster@ua.edu

Jake Yue Chen
School of Medicine
University of Alabama at Birmingham
jakechen@uab.edu

Ying Ding
School of Information
University of Texas at Austin
ying.ding@ischool.utexas.edu

Abstract

We present an interactive visualization of the Cell Map for AI Talent Knowledge Graph (CM4AI TKG), a detailed semantic space comprising approximately 28,000 experts and 1,000 datasets focused on the biomedical field. Our tool leverages transformer-based embeddings, WebGL visualization techniques, and generative AI, specifically Large Language Models (LLMs), to provide a responsive and user-friendly interface. This visualization supports the exploration of around 29,000 nodes, assisting users in identifying potential collaborators and dataset users within the health and biomedical research fields. Our solution transcends the limitations of conventional graph visualization tools like Gephi, particularly in handling large-scale interactive graphs. We utilize GPT-4o to furnish detailed justifications for recommended collaborators and dataset users, promoting informed decision-making. Key functionalities include responsive search and exploration, as well as GenAI-driven recommendations, all contributing to a nuanced representation of the convergence between biomedical and AI research landscapes. In addition to benefiting the Bridge2AI and CM4AI communities, this adaptable visualization framework can be extended to other biomedical knowledge graphs, fostering advancements in medical AI and healthcare innovation through improved user interaction and data exploration. The demonstration is available at: https://jiawei-alpha.vercel.app/.

1 Introduction

This paper presents an interactive visualization of the Bridge2AI - Cell Maps for AI Data Generation Project Talent Knowledge Graph (CM4AI TKG) [2, 18] semantic space. The Bridge2AI project [14] assembles professionals from biomedicine, computer science, and social sciences to develop and curate high-quality biomedical datasets critical for AI-driven research and transformative healthcare

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

advancements. The CM4AI TKG is designed to assist in identifying suitable potential users for these datasets [11] and locating relevant collaborators within the extensive biomedical domain. Our visualization tool supports the exploration of the distribution of CM4AI researchers and datasets within the broader biomedical and genomics landscape, thereby aiding the discovery of future collaborators and users. We use Large Language Models (LLM) to offer suggestions for potential collaborators and users for each researcher and biomedical dataset in the CM4AI TKG.

Our visualization application is highly adaptable, incorporating node representations via transformer-based semantic embeddings [13], dimensionality reduction techniques like t-SNE and UMAP [16, 8], and large-scale responsive node visualization through PixiJS [17]. This configuration allows for the future integration of additional functionalities. Compared to widely used graph visualization tools, such as Gephi [1] and Cytoscape [12], PixiJS efficiently handles and visualizes large datasets while offering excellent cross-platform capabilities. Furthermore, we incorporate LLM to enhance user understanding and provide explainability. This interactive visualization framework can be adapted for other medical domain knowledge graphs, offering an interactive interface that improves user accessibility.

The CM4AI TKG was extracted from the PubMed Knowledge Graph [19] and Semantic Scholar [7]. It includes 2 million papers, 44,000 authors, 1,179 biomedical datasets [11] used in these studies, and bio-entities mentioned in the papers. The interactive visualization of the CM4AI TKG provides an intuitive, user-friendly interface for exploring the knowledge graph. Utilizing semantic information from paper titles and abstracts [13], our system maps a two-dimensional semantic space, indicating the relative positions of researchers and datasets within the CM4AI knowledge domain. Distances between researchers and datasets reflect their similarities. Users can search for datasets or talents, navigate the space with zoom functionality, and access detailed author and dataset information.

The following sections cover the data and methodology used in constructing the CM4AI TKG, along with an outline of its functionalities and practical applications.

2 Data and Methods

2.1 Data Preparation

The CM4AI TKG compiles data related to 121 core researchers participating in the Bridge2AI project, including 35 researchers identified as core members of CM4AI. This dataset encompasses these researchers' 44,000 co-authors and approximately 2.05 million publications in total. Out of these, we specifically employed 10,011 publications associated directly with the core researchers.

To construct the dataset, we first collected ORCID [4] identifiers for the 121 core researchers and manually confirmed their identities using Semantic Scholar to ensure accuracy. Following this, we matched these researchers within the PubMed Knowledge Graph [19] to gather comprehensive data on their publications, co-authors, datasets used, and mentioned biological entities. To enhance relevance and focus, we excluded any authors without publications after 2020, refining our dataset to comprise 28,000 active researchers, referred to as 'talents,' for visualization purposes.

2.2 Author and Dataset Representation

To represent authors and datasets, we employed Specter2, a state-of-the-art BERT-based encoder [13, 6], to transform the titles and abstracts of all 2.05 million papers into 768-dimensional embedding vectors. For author representation, we aggregated these embeddings by weighing the author's position in each publication. Specifically, the first and last authors received a weight of 1, while a k-th author was assigned a weight of $\frac{1}{k}$. Authors beyond the 10th position were assigned a uniform weight of $\frac{1}{10}$. Dataset representations were obtained by aggregating embeddings from papers that utilized the respective datasets. To identify potential collaborators, we selected the top 30 researchers for each author—individuals with whom they had never collaborated—by computing cosine similarity of the embeddings. Similarly, for each dataset, we identified the top 150 researchers who had not used the dataset as potential users.

We utilized GPT-4o [9] to provide justifications for these recommendations. For collaborator recommendations, we inputted to the model five recent and five most-cited papers since 2017 for each author, along with metadata such as title, journal, citation count, and publication year [5]. The model then

generated justifications highlighting the potential benefits of collaboration. For recommending dataset users, we provided the model with an author's recent and highly cited papers and accompanying dataset descriptions, prompting the model to justify why these users should consider the dataset.

2.3 Visualization

For visualizing the data, we utilized PixiJS [17], a versatile 2D WebGL renderer, capable of efficiently displaying large numbers of nodes in a two-dimensional space. Unlike traditional graph visualization tools such as Gephi [1] and Cytoscape [12], PixiJS takes advantage of WebGL technology to handle and render large datasets, ensuring high cross-platform compatibility. This allows users to explore the visualization effortlessly through any modern web browser. The web application was developed using TypeScript and Svelte, with visualization and hosting code adapted from an open-source anime recommendation project, Sprout [10]. Additionally, an Oracle APEX visualization is available, providing detailed information such as each talent's publication history, accessible at: https://cm4ai.org/ckg.

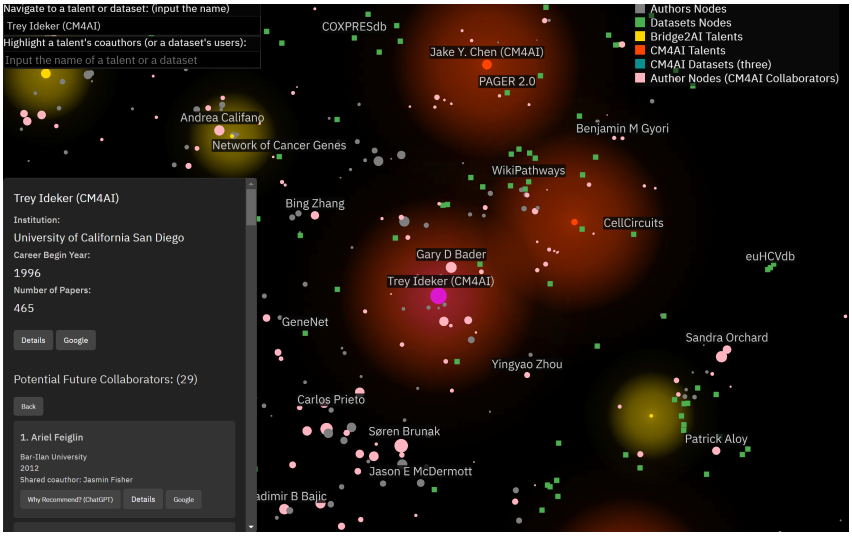
To ensure a well-organized layout, crucial for effective embedding visualization and user experience, we experimented with various dimensionality reduction techniques and layout configurations using Emblaze [3]. This tool facilitates the visual comparison of embedding spaces and offers built-in dimensionality reduction methods like t-SNE and UMAP [16, 8]. These methods were used to condense the 768-dimensional embeddings of authors and datasets into two-dimensional coordinates. We tuned the parameters of t-SNE and UMAP to achieve an optimal layout that enhances visualization quality and user interaction.

3 Function and Use Cases

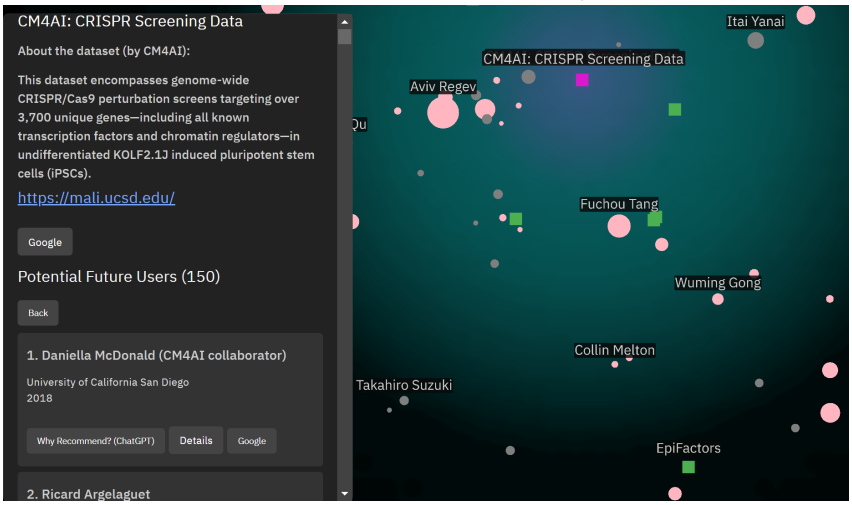
The CM4AI TKG semantic space visualizes talents and datasets as nodes within a two-dimensional space, as illustrated in Figure 1. Node size reflects the number of publications linked to each talent (log scale), while node shapes and colors distinguish between node types: dataset nodes are depicted as squares and talent nodes as circles. This interactive interface allows users to drag, scroll, zoom, hover over, and search nodes across approximately 28,000 talent nodes and 1,179 dataset nodes. The system provides two main functionalities: (1) Exploration of existing positions in the knowledge space and (2) Recommendation of potential collaborators or dataset users with justifications provided by GPT-4o.

Explore Existing Position in the Knowledge Space. Users can efficiently locate and explore a specific dataset or talent by entering its name in the search box. The system then provides a list of candidates; once selected, such as 'Trey Ideker' shown in Figure 1a, the view zooms to this talent, presenting an information window with key details such as institution, number of publications, and career start year. By clicking the 'Detail' button, users are directed to an Oracle APEX interface to review detailed publication histories. Similarly, searching for a dataset reveals detailed information in a popup window, as demonstrated with 'CRISPR Screening Data' in Figure 1b. The 'Explore Existing Collaborators or Users' feature helps users understand current collaboration and usage patterns. By entering a talent's name, such as 'Trey Ideker', in the second search box, users can highlight the nodes of talents who have previously collaborated with that individual, creating a visually distinct starry effect (Figure 2a). This facilitates comparisons between historical and potential future collaborations.

Use of Generative AI for Informed Recommendations. The system also provides lists of recommended collaborators or dataset users, accompanied by LLM-generated justifications explaining why these individuals are suggested. By clicking the 'Why Recommend?' button, users can view these justifications. Users can explore these potential collaborators by clicking their names, redirecting the view to their respective nodes. This feature leverages the reasoning capabilities of LLMs to enhance the usefulness of recommendations. In Figure 2a, for instance, the model justifies recommending Ariel Feiglin to Trey Ideker, and in Figure 2b, it provides reasons for suggesting Daniella MacDonald as a user for the CRISPR Screening Data.



(a) Information Window for a Talent (Trey Ideker)

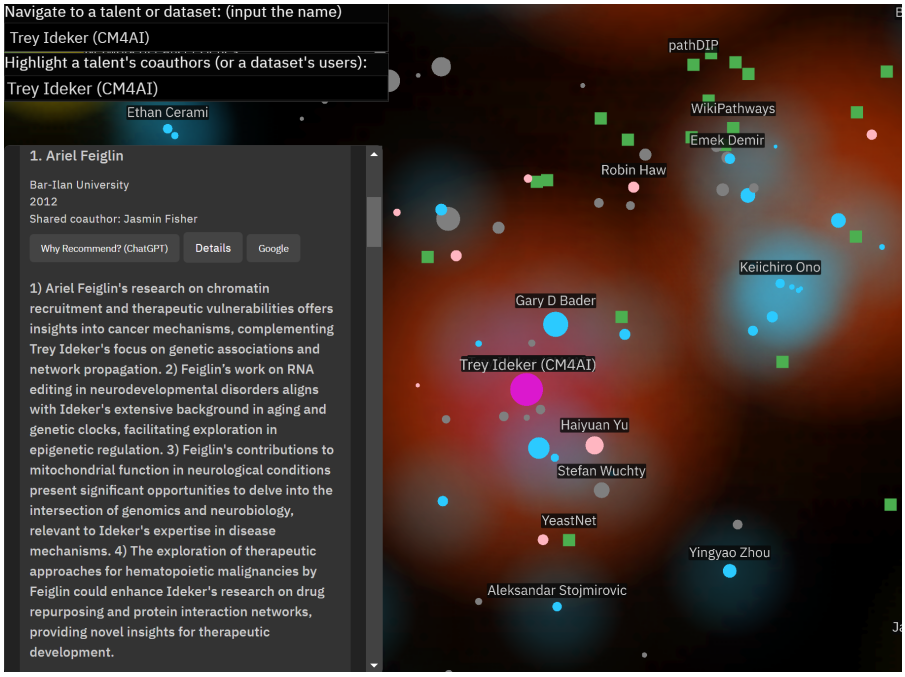


(b) Information Window for a Dataset (CRISPR Screening Data)

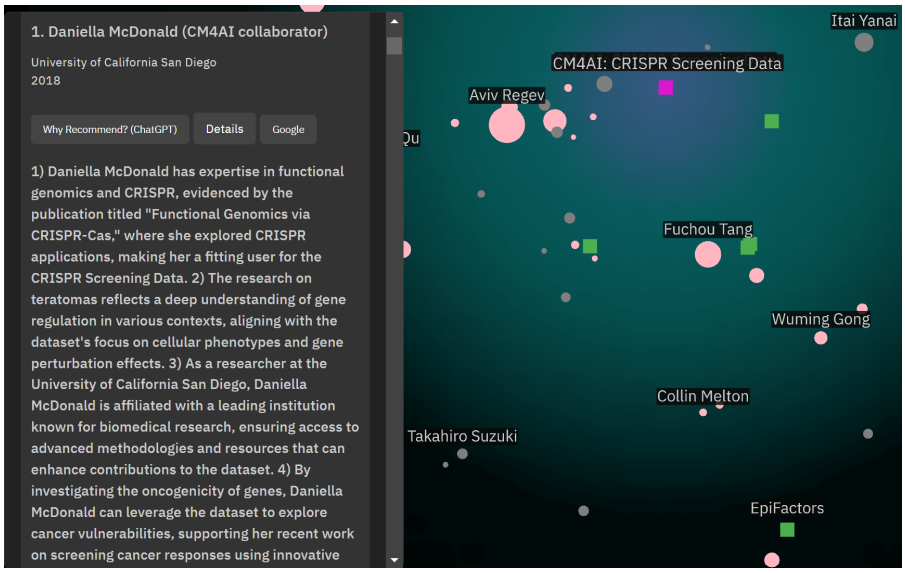
Figure 1: Information windows for different items: (a) a talent, (b) a dataset.

4 Summary

The current demonstration offers a rapid-response interface for users to interact with the CM4AI Talent Knowledge Graph (TKG). It integrates user recommendations with explanations generated by LLMs that are based on detailed background information. Additionally, it displays profiles of experts and datasets with further details via both an information window and a database visualization. The semantic space of the CM4AI TKG provides an intuitive and user-friendly interface, which showcases the growth of the Bridge2AI and CM4AI community within the expansive biomedical landscape. This WebGL-based visualization technology is not only advantageous for the Bridge2AI and CM4AI communities but can also be adapted for other domain-specific knowledge graphs containing hundreds of thousands of nodes. There are some limitations to be addressed in this work: (1) Author name disambiguation and normalization [15]. The user experience can be affected by inconsistencies in author representation. Some authors may not be accurately disambiguated, leading to discrepancies, such as variations in name representation, where some authors are reported with full names while others only have initials. (2) Subjective evaluation of visualization. The evaluation of the final visualization is relatively subjective. Using dimensionality reduction methods like t-SNE or UMAP to determine node coordinates, the current assessment is based on the visual clustering



(a) LLM's Justification for Recommending Ariel Feiglin to Trey Ideker



(b) LLM's Justification for Recommending Daniella MacDonald to CRISPR Screening Data

Figure 2: LLM's Justifications for Recommendations: (a) For Trey Ideker, (b) For CRISPR Screening Data

and interpretability of node positions. Future work should focus on developing objective methods to evaluate the visualization's effectiveness.

Overall, this visualization pipeline provides considerable advantages for medical applications that require visualizing extensive datasets and customizing frameworks to incorporate LLM reasoning. By continually refining these capabilities, we can enhance the utility of large-scale knowledge graphs or vector (embedding) data. This advancement drives progress in medical AI and research by offering user-friendly information representation, thereby facilitating better data interpretation and decision-making in healthcare innovations.

Acknowledgments

We would like to acknowledge the following funding supports: NIH 1OT2OD032742-01, NIH OTA-21-008, NIH OT2OD032581, NSF 2333703, NSF 2303038.

References

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

[2] T. Clark, J. Mohan, L. Schaffer, K. Obernier, S. Al Manir, C. P. Churas, A. Dailamy, Y. Doctor, A. Forget, J. N. Hansen, et al. Cell maps for artificial intelligence: Ai-ready maps of human cell architecture from disease-relevant cell lines. *bioRxiv*, 2024.

[3] C. D. I. Group. cmudig/emblaze, Aug. 2024. URL `https://github.com/cmudig/emblaze`. original-date: 2021-08-13T18:57:40Z.

[4] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner. Orcid: a system to uniquely identify researchers. *Learned publishing*, 25(4):259–264, 2012.

[5] B. I. Hutchins, K. L. Baker, M. T. Davis, M. A. Diwersy, E. Haque, R. M. Harriman, T. A. Hoppe, S. A. Leicht, P. Meyer, and G. M. Santangelo. The nih open citation collection: A public access, broad coverage resource. *PLoS biology*, 17(10):e3000385, 2019.

[6] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[7] R. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

[8] L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018. URL `https://api.semanticscholar.org/CorpusID:3641284`.

[9] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[10] C. Primozic. Ameobea/sprout, Sept. 2024. URL `https://github.com/Ameobea/sprout`. original-date: 2022-04-18T20:34:02Z.

[11] D. J. Rigden and X. M. Fernández. The 2023 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 51(D1):D1–D8, 2023.

[12] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[13] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL `https://api.semanticscholar.org/CorpusID:254018137`.

[14] M. Suran. New nih program for artificial intelligence in research. *JAMA*, 328(16):1580–1580, 2022.

[15] V. I. Torvik and N. R. Smalheiser. Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3):1–29, 2009.

[16] L. van der Maaten and G. E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `https://api.semanticscholar.org/CorpusID:5855042`.

[17] R. Van der Spuy. *Learn Pixi. js*. Apress, 2015.

[18] H. Xu, C. Gupta, Z. Sembay, S. Thaker, P. Payne-Foster, J. Chen, and Y. Ding. Cross-team collaboration and diversity in the bridge2ai project. In *Companion Proceedings of the ACM Web Conference 2023*, pages 790–794, 2023.

[19] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, Y. Bu, C. Chen, I. A. Ebeid, D. Li, and Y. Ding. Building a PubMed knowledge graph. *Scientific Data*, 7(1):205, June 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0543-2. URL `https://www.nature.com/articles/s41597-020-0543-2`.