Research and Patent Clusters Dataset: Charting the World's Open Research and Innovation Landscape

Rebecca Gelles[1], Katherine Quinn[1], and Catherine Aiken[1]

[1]Center for Security and Emerging Technology, Georgetown University, Washington D.C., United States, 22152

September 2026

1 Introduction

The rapid pace of scientific and technological development creates informational challenges for researchers and decision makers. Meanwhile, there is a lack of use-inspired, accessible open datasets and tools for self-serve analysis of worldwide scientific and patenting landscapes. We fill this gap by providing an open research and patent clusters dataset. We aggregate global scientific research and patent documents into clusters using citations and semantic similarity [Janssens et al., 2009, Boyack and Klavans, 2020, Thijs et al., 2015], identifying granular areas of research and patenting.

To be useful for analysis and decision-making, we curate original cluster-level metadata to characterize the research and patent clusters, with a focus on metrics to support the identification of emerging or high-impact areas [Boyack et al., 2014, Upham and Small, 2010, Small et al., 2014, Weis and Jacobson, 2021, Krenn et al., 2023] and application of science [Lewis et al., 2024, Schmallenbach et al., 2024, Heldens et al., 2020, Jeong et al., 2022]. Specifically, we augment the cluster dataset with original metadata to identify recent research activity, predict cluster growth, summarize and classify cluster contents, flag key entities, and connect research and patent activity.

We present three example research uses of the cluster dataset. The first and second explores federally-funded research, through an exploration of Research Diffusion Pathways and network effects. The third traces the development of GLP-1 Agonists from research to innovation.

The research and patent cluster dataset is available on Zenodo (https://doi.org/10.5281/zenodo.19665550) and is updated monthly. Our clusters can also be explored through CSET's Map of Science (https://sciencemap.eto.tech) and Map of Patents (https://patentmap.eto.tech).

2 Dataset Construction

We combine and de-duplicate several datasets to generate a corpus of scientific literature and patents, which we use for clustering. Our goal is to obtain global coverage of all openly available data with accurate metadata. Our research corpus combines four databases of research literature: OpenAlex[1], Semantic Scholar[2], The Lens[3], and arXiv.[4] We drop papers with missing titles and only include papers with at least one citation or reference, resulting in just over 146 million research documents (in 163 different languages) for clustering.

Our patents corpus combines data from The Lens [5], EPO's PATSTAT[6], and 1790 Analytics[7]. We use

[1]https://openalex.org/
[2]https://www.semanticscholar.org/
[3]https://about.lens.org/
[4]https://arxiv.org/
[5]https://about.lens.org/
[6]https://www.epo.org/en/about-us/observatory-patents-and-technology/observatory-tools/patstat
[7]https://1790analytics.com/

PATSTAT's simple family id to de-duplicate inventions and apply quality and type limitations, resulting in 94 million patent families that come from 107 different patent authorities.

Figure 1 displays the steps in developing our research and patent clusters, extracting metadata, and visualizing them in open tools.

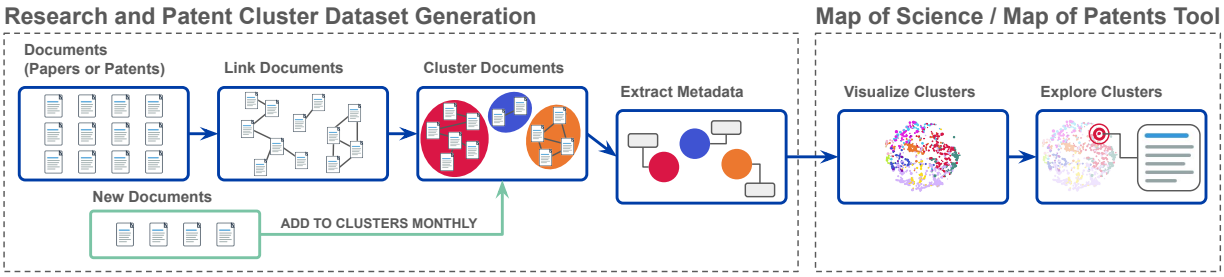


Figure 1: Flow diagram of cluster generation and visualization in Map of Science and Map of Patents.

First, we generate a hybrid network for each set of documents using citations and text similarity [Boyack and Klavans, 2020, Quinn et al., 2026b, 2025b]. For research, this includes citations and text similarity, based on embedding title and abstract text using a pretrained sentence transformer model and calculating the cosine similarity between them. For patents we also embed the English text of their Cooperative Patent Classification (CPC) or International Patent Classification (IPC) codes to assess text similarity across documents. To maintain a reasonable number of network edges and avoid prohibitive computational costs, we find the top ten most similar documents for each document in our dataset using FAISS [Johnson et al., 2019].

Next, we apply the Leiden clustering algorithm [Traag et al., 2019] on the hybrid research and patent networks.[8] We aim for about 100 thousand clusters in each document set, a number which provides a useful level of granularity for finding emerging areas [Rahkovsky et al., 2021]. We obtain nearly 92 thousand research clusters and over 116 thousand patent clusters. These resulting clusters are the central unit of analysis for our dataset.

2.1 Developing Cluster Metadata

To augment the dataset, we extract original metadata for each cluster. This metadata enables cluster searching, sorting, filtering, and description. We attempted, where possible, to make research and patent cluster metadata interoperable for improved linkage and ease of use. Curated original cluster metadata includes:

LLM Extreme Summarization: We generate descriptive titles (5-7 words) and summaries (15-20 words) for each cluster [Le et al., 2026] using Gemini. Outputs are validated by subject matter experts to check for hallucinations and ensure accurate interpretation.

Topic Classification: We assign fields of study to our research clusters using a taxonomy of 1,100 fields over four hierarchical levels [Gelles and Dunham, 2024]. We also classify research clusters according to their relevance to AI, natural language processing, robotics, computer vision, cybersecurity, AI safety, LLM development, and chip design and fabrication [Toney-Wails et al., 2024, Dunham et al., 2020, Aiken et al., 2024]. We categorize patent clusters, based on CPC/IPC codes, using topics derived from the United Nation's International Standard of Industrial Classification of all Economic Activities (ISIC) [Department of Economic and Social Affairs, 2008]. We also categorize patent cluster into five top-level emerging technology categories: AI, robotics, cybersecurity, biotechnology, and semiconductors, and various subcategories based on the ACM Computing Classification System [Rous, 2012], the IEEE Robots taxonomy [Isaka, 2024], and subject matter expert input.

Entity Resolution and Matching: For research clusters, we provide top affiliated authors, organizations, and countries. For organizations and countries, we run our own entity resolution pipeline and link entities to the Research Organization Registry (ROR)[9] where available. For patent clusters, we provide top countries, assignees, and inventors. Ongoing work includes text-based improvement of assignee data.

[8] Details on clustering parameters, iterations and optimization modularity are provided in Le et al. [2026]

[9] https://ror.org/

Annual Production: We provide an annual count of documents in each cluster based on publication dates. For patents, we consider the number of new patent families filed in that year for annual counts. We further break down yearly counts by country, as well as cross-country collaborations (for papers) and cross-jurisdiction filings (for patents).

Growth: We determine current growth ratings based on recent activity, ranking clusters by percentile. For research clusters this is determined by publication dates; for patents we use grant dates. We also classify research clusters as *likely*/*not likely* to experience extreme growth in the next three years [Quinn and Toney-Wails, 2026, Quinn, 2026c, Klavans et al., 2020, Rahkovsky et al., 2021].[10]

3 Dataset Uses

To demonstrate possible uses of this dataset, we share initial results from three ongoing analyses. These examples do not represent the full range of uses of the clusters dataset.

3.1 Research Diffusion Pathways and Tracing NIH Funding

In Toney-Wails et al. [2026], we construct Research Diffusion Pathways (RDPs), defined as sequences of research clusters connected through inter-cluster citations. Citation links can exist between clusters (via publications that cite or are cited by other work outside of their cluster) and connect publications across distinct areas of science. We refer to these inter-cluster connections as *diffusion points*, which provide observable instances in which research is cited outside the research area to which it is most closely connected. We summarize these *diffusion points* using Gemini 2.5 Flash to generate *link summaries*, which are themselves summarized to create Citation Aware Research Diffusion Summaries (CARDS). We show an example RDP and its overall CARDS in Figure 2.

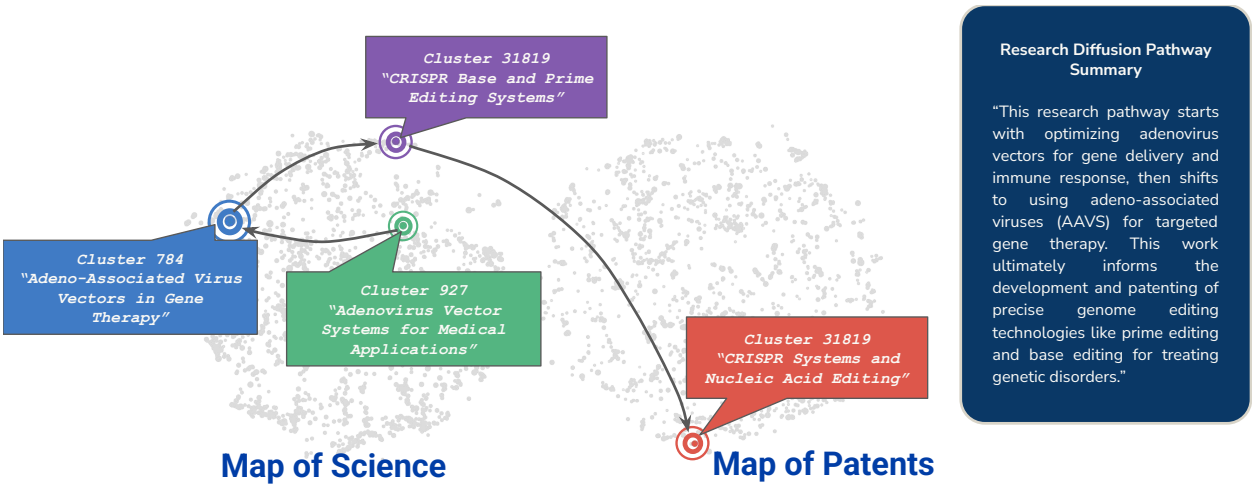


Figure 2: Example research diffusion pathway consisting of four cluster connected through three diffusion points, starting in the Map of Science and ending in the Map of Patents, with overall LLM-generated summary.

We construct RDPs and summarize them with CARDS to better understand diffusion through the landscape of federally-funded research. We study the National Institutes of Health (NIH) by finding papers the NIH RePORTER database[11] ($N = 3$ million papers) and matching them to their assigned research clusters ($N = 50$ thousand clusters). We identify top clusters with NIH funded research, defined as the top

[10]The classifier is a fully connected linear neural network with one hidden layer, which incorporates 141 features of interest from each research cluster. A full list of cluster features, as well as neural network details and code are provided on GitHub at https://github.com/georgetown-cset/research_cluster_extreme_growth_classifier.

[11]https://reporter.nih.gov/

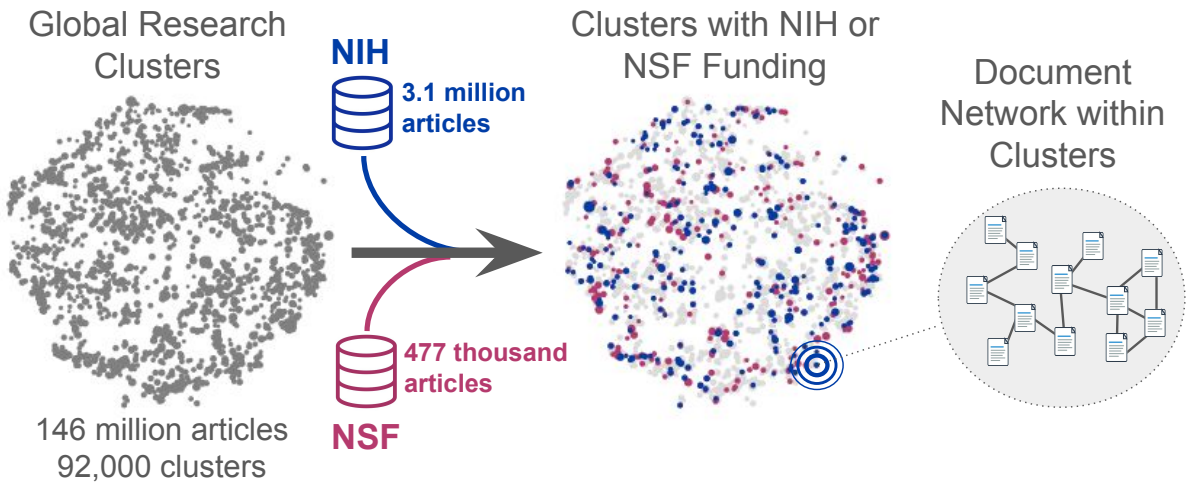


Figure 3: We match articles with NIH or NSF funding to their corresponding research clusters within the Map of Science. Since each cluster also contains a network of connected scientific documents, we can further explore these network to understand the impact of these agencies' funding on areas of science.

10% of NIH-funded clusters by concentration (resulting in a set of 4,826 clusters). We construct RDPs with four diffusion points, and categorize them by field.

As expected given the NIH's focus on biomedical research, initial clusters are predominantly classified as biology (78.6%). However, after following the RDPs through four diffusion points, the proportion of biology clusters has decreased to 59.5%, while medicine grew from 1.9% to 13.4%, social science from 16.2% to 19.4%, materials science from 0.4% to 2.8%, and computer science from 0.3% to 1.7%. RDPs become more diffuse throughout the topical landscape, consistent with knowledge spreading from basic research (biology) to applied (medicine).

3.2 Structural Role of the NIH and NSF in Science Networks

In Quinn and Toney-Wails [2026], we apply network analyses to understand the structural role of research funded by the NIH and the National Science Foundation (NSF) in different areas of science. We use the same NIH RePORTER database described in Section 3.1 to find 3.1 million NIH-funded publications and funding data from the Lens to identify over 477,000 articles with NSF funding (Figure 3).

Using the Rényi and Erdos [1959] and Gilbert [1959] model of random graphs [Fienberg, 2012] applied to the hybrid networks in each cluster, we compare the network weight of edges connected to funded papers. We find NIH-funded publications have observed contribution to network edge weight equivalent to a research portfolio that is 45.8% larger. The corresponding effect is 42.3% for NSF-funded publications and 45.5% when considering publications funded by either NIH or NSF.

We generate funding-shock simulations by evaluating five funding shock scenarios (10%, 20%, 30%, 40%, and 50%), representing incremental increases in the proportion of NIH- or NSF-funded publications removed from each cluster. For each shock level, we perform multiple random iterations, removing the specified proportion of funded publications and, as a baseline comparison, an equal number of randomly sampled non-funded publications. We compare the network weighting lost in each scenario.

Overall, NIH-funded publication removal results in a 76.2% greater loss in *incoming* edge weight than matched non-funded publication removal, compared with 31.5% for NSF funded publications and 73.8% when considering NIH or NSF funded publications together. Differences in *out-degree* edge weight are considerably smaller, with relative increases of 3.6% (NIH), 14.5% (NSF), and 3.9% (NIH or NSF). The observed stronger effect on in-degree edge weight loss indicates that funded publications (notably those supported by the NIH) receive stronger connections from other publications in their research networks, suggesting that subsequent

Table 1: Research clusters most frequently cited by GLP-1 agonist patent cluster

Cluster	Citations	Title
11442	258	Computational approaches to protein structure prediction
10073	110	GLP-1 Receptor Agonists for Weight Management
1399	38	Glucagon-like Peptide-1 Receptor Signaling Mechanisms
10381	30	Glucose-Dependent Insulinotropic Polypeptide Metabolic Functions
5991	27	Sustainable peptide synthesis processes
49508	16	Design and Application of Therapeutic Peptides
58875	14	Phylogenetic Tree Visualization and Annotation Tools
3371	11	Glucagon-like Peptide Receptor Agonist Therapies
4141	11	Lysophosphatidic Acid Signaling and Biological Functions

research disproportionately connects to and builds upon federally funded work.

3.3 A Case Study: Ozempic

In Quinn et al. [2025a], we trace the research leading to the discovery and commercialization of the blockbuster GLP-1 Agonist drugs (Ozempic, Wegovy, Mounjaro, and Zepbound). The active ingredients in GLP-1 agonist drugs, held by companies such as Eli Lilly, Hanmi Pharmaceuticals[12], and Novo Nordisk, are all located in the same patent cluster, cluster 2081: "Methods for synthesizing and developing peptide compounds acting as dual receptor agonists for potential metabolic health applications."[13]

Using patent-to-paper citations, we link the GLP-1 agonist patent cluster to underlying research clusters.[14] We find 3,169 research clusters (about 3% of all research clusters) are linked through direct citations to the GLP-1 Angoist patent cluster. Nine clusters are cited at least 10 times, as seen in Table 1, and include research areas related to the drug's development and application, underlying pharmaceutical processes, and adjacent and new growth areas. The GLP-1 Agonist cluster is not an an anomaly. We find that, on average, the 7,010 patent clusters related to Life Sciences each draw on 281 research clusters, consistent with innovations drawing on multiple sources rather than being the direct product of a single, linear process.

4 Dissemination and Next Steps

The research and clusters dataset is a public dataset with global coverage and use-inspired metadata. It includes 92 thousand research clusters and 116 thousand patent clusters and is easily accessible on Zenodo (https://doi.org/10.5281/zenodo.19665550) and available for use under CC BY 4.0.[15] As an alternative way to engage with the data, we developed and maintain two publicly available tools, the Map of Science (https://sciencemap.eto.tech) and the Map of Patents (https://patentmap.eto.tech). We also highlight research uses, new features, and quick analysis enabled by the dataset in blog posts and data snapshots [Quinn et al., 2025b,a, Quinn, 2026c,b,a, Le, 2026, Quinn et al., 2026a]. CSET is committed to maintaining the dataset and related resources as open research resources.

There are numerous opportunities for expanded use and exploration of the dataset. As seen in our case studies, exploring the landscape of research and innovation can provide significant insights on the impacts of research funding and the trajectories to groundbreaking innovation. Beyond that we believe the dataset can bolster portfolio analysis, help monitor new developments and activity, and inform research agenda and strategy-setting. The dataset likely has use cases we have not thought of or explored. We hope to learn more about those possibilities in this meeting!

[12] Hanmi Pharmaceuticals is currently preparing to launch South Korea's first domestic GLP-1 drug [Hyun-woo, 2026].

[13] More information about the GLP-1 agonist patent cluster can be found here: https://patentmap.eto.tech/v1/cluster/2081

[14] Although patent-to-paper links are not yet available in the open clusters, we can provide these connections on request.

[15] Additional data for paper assignment to clusters and mapping OpenAlex ids to research clusters is also published on Zenodo (https://zenodo.org/doi/10.5281/zenodo.11034260)

References

Catherine Aiken, James Dunham, Jennifer Melot, and Zachary Arnold. Identifying emerging technologies in research. *Center for Security and Emerging Technology*, 2024. doi: 10.51593/20240050.

Kevin W Boyack and Richard Klavans. A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, 1(4):1570–1585, 2020. doi: 10.1162/qss_a_00085.

Kevin W Boyack, Richard Klavans, Henry Small, and Lyle Ungar. Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management*, 32:147–159, 2014.

Department of Economic and Social Affairs. International standard industrial classification of all economnic activities, revision 4. Statistical papers ST/ESA/STAT/SER.M/4/Rev.4, United Nations Statistics Division, 2008. URL https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4e.pdf.

James Dunham, Jennifer Melot, and Dewey Murdick. Identifying the development and application of artificial intelligence in scientific text. *arXiv preprint arXiv:2002.07143*, 2020. doi: 10.48550/arXiv.2002.07143.

Stephen E. Fienberg. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, 2012. doi: 10.1080/10618600.2012.738106.

Rebecca Gelles and James Dunham. Multi-label field classification for scientific documents using expert and crowd-sourced knowledge. In Lucie Lucie-Aimée, Angela Fan, Tajuddeen Gwadabe, Isaac Johnson, Fabio Petroni, and Daniel van Strien, editors, *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 14–20, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wikinlp-1.7. URL https://aclanthology.org/2024.wikinlp-1.7/.

Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Stijn Heldens, Pieter Hijma, Ben Van Werkhoven, Jason Maassen, Adam SZ Belloum, and Rob V Van Nieuwpoort. The landscape of exascale research: A data-driven literature analysis. *ACM Computing Surveys (CSUR)*, 53(2):1–43, 2020.

Nam Hyun-woo. Hanmi pharm in spotlight as korea's 1st obesity drug nears launch. *The Korea Times*, August 2026. URL https://www.koreatimes.co.kr/business/companies/20260826/hanmi-pharm-in-spotlight-as-koreas-1st-obesity-drug-nears-launch.

Satoru Isaka. A taxonomic classification and identification system for robots: Abstract. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3799–3800, 2024. doi: 10.1109/SMC54092.2024.10831650.

Frizo Janssens, Lin Zhang, Bart De Moor, and Wolfgang Glänzel. Hybrid clustering for validation and improvement of subject-classification schemes. *Information processing & management*, 45(6):683–702, 2009.

Jongkil Jay Jeong, Yevhen Zolotavkin, and Robin Doss. Examining the current status and emerging trends in continuous authentication technologies through citation network analysis. *ACM Computing Surveys*, 55(6):1–31, 2022.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572.

Richard Klavans, Kevin W Boyack, and Dewey A Murdick. A novel approach to predicting exceptional growth in research. *Plos one*, 15(9):e0239177, 2020. doi: 10.1371/journal.pone.0239177.

Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P Moutinho, Nima Sanjabi, et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5(11):1326–1335, 2023.

Jack Le. Hot research topics in ai and cybersecurity: insights from the map of science, 2026. URL https://eto.tech/blog/hot-topics-cybersecurity-ai-map-of-science/.

Jack Le, Rebecca Gelles, Catherine Aiken, and Katherine Quinn. Extreme summarization of worldwide scientific and patent document clusters for landscape analysis. Manuscript under review, 2026.

Joanna I Lewis, Autumn Toney, and Xinglan Shi. Climate change and artificial intelligence: assessing the global research landscape. *Discover Artificial Intelligence*, 4(1):64, 2024.

Katherine Quinn. Hot topics in ai and bioinformatics: insights from the map of science, 2026a. URL https://eto.tech/blog/hot-topics-ai-bioinformatics-map-of-science/.

Katherine Quinn. Canadian funding agencies' impact on the research and innovation landscape, 2026b. URL eto.tech/blog/canadian-funding-agencies-impact-research-innovation-landscape/.

Katherine Quinn. Predicting areas of extreme growth with the map of science, 2026c. URL https://eto.tech/blog/predicting-areas-of-extreme-growth-with-map-of-science/.

Katherine Quinn and Autumn Toney-Wails. Measuring the outsized structural role of u.s. federally funded research in science networks. Manuscript under review, 2026.

Katherine Quinn, Steph Batalis, and Gelles. The nih's impact on research and innovation, 2025a. URL https://cset.georgetown.edu/publication/the-nihs-impact-on-research-and-innovation/.

Katherine Quinn, Rebecca Gelles, Ronnie Kinoshita, and Christian Schoeberl. Introducing cset's patent clusters, 2025b. URL https://cset.georgetown.edu/publication/introducing-csets-patent-clusters/.

Katherine Quinn, Steph Batalis, and Gelles. Artificial intelligence patent clusters, 2026a. URL https://cset.georgetown.edu/publication/artificial-intelligence-patent-clusters/.

Katherine Quinn, Rebecca Gelles, Adrian Thinnyun, Brian Love, and Catherine Aiken. Map of science: Charting worldwide open scientific literature. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*. ACM, 2026b. doi: 10.1145/3799682.3840244. Accepted for publication.

Ilya Rahkovsky, Autumn Toney, Kevin W Boyack, Richard Klavans, and Dewey A Murdick. Ai research funding portfolios and extreme growth. *Frontiers in Research Metrics and Analytics*, 6:630124, 2021. doi: 10.3389/frma.2021.630124.

Alfréd Rényi and Paul Erdos. On random graphs. *Publicationes Mathematicae*, 6(290-297):5, 1959.

Bernard Rous. Major update to acm's computing classification system. *Communications of the ACM*, 55 (11):12–12, 2012. doi: 10.1145/2366316.2366320.

Leo Schmallenbach, Till W Bärnighausen, and Marc J Lerchenmueller. The global geography of artificial intelligence in life science research. *Nature Communications*, 15(1):7527, 2024.

Henry Small, Kevin W Boyack, and Richard Klavans. Identifying emerging topics in science and technology. *Research policy*, 43(8):1450–1467, 2014.

Bart Thijs, Wolfgang Glänzel, and Martin S Meyer. Using noun phrases extraction for the improvement of hybrid clustering with text-and citation-based components. the example of "information systems research". In *Proc. of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey*, volume 1384, pages 28–33, 2015.

Autumn Toney-Wails, Christian Schoeberl, and James Dunham. Ai on ai: Exploring the utility of gpt as an expert annotator of ai publications. *arXiv preprint arXiv:2403.09097*, 2024. doi: 10.48550/arXiv.2403.09097.

Autumn Toney-Wails, Steph Batalis, and Katherine Quinn. Cards: Citation-aware research diffusion summarization. Under review, 2026.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019. doi: https://doi.org/10.1038/s41598-019-41695-z.

S Upham and Henry Small. Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83(1):15–38, 2010.

James W Weis and Joseph M Jacobson. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, 39(10):1300–1307, 2021.