Patterns of Gender and Funding Diversity in post-2022 publications on "LLM" and "ChatGPT"

Moses Boudourides

Master's in Data Science Online Program
School of Professional Studies
Northwestern University

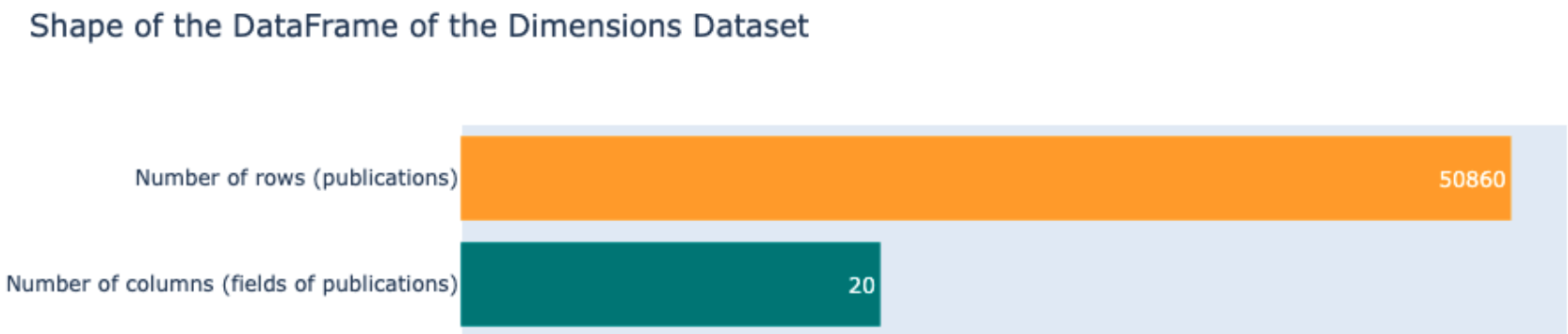
Moses.Boudourides@northwestern.com

TUM Think Tank
School of Social Sciences and Technology
Technische Universität München
Thursday, October 31, 2024

A Keywords Search from Dimensions.ai

- **Why choose Dimensions over Scopus, Web of Science, or another bibliometric database?** My primary reasons in selecting Dimensions were twofold:
  (i) it provides a high degree of completeness and quality in publication metadata (Delgado-Quirós, 2024; Nguyen et al., 2022), and
  (ii) it offers a Python client, `dimcli`, which facilitates programmatic access to the Dimensions API for efficient querying (`https://api-lab.dimensions.ai/cookbooks/1-getting-started/5-Deep-dive-DSL-language.html`).
- The dimensions query was "`%dslloopdf search publications in title_abstract_only for 'chatgpt' or 'large language model' or 'LLM' where year = 2022-2025 return publications`."
- The total number of attributes of publications and grants were 20 fields, among which we are analyzing in this study the following 9: `id`, `authors`, `title`, `date`, `doi`, `type`, `category_for`, `grants` `funding_usd` (definitions: `https://docs.dimensions.ai/dsl/2.0.0/datasource-publications.html`).

After removing duplicate publications, the unique count of each publication in the collected Dimensions dataset was established by enumerating the distinct `id` field associated with each publication. It is important to note that if a publication was retrieved under two or more different `types`—for example, as both an `article` (or `proceeding` or `chapter`) and a `preprint`—the `preprint` version was excluded.

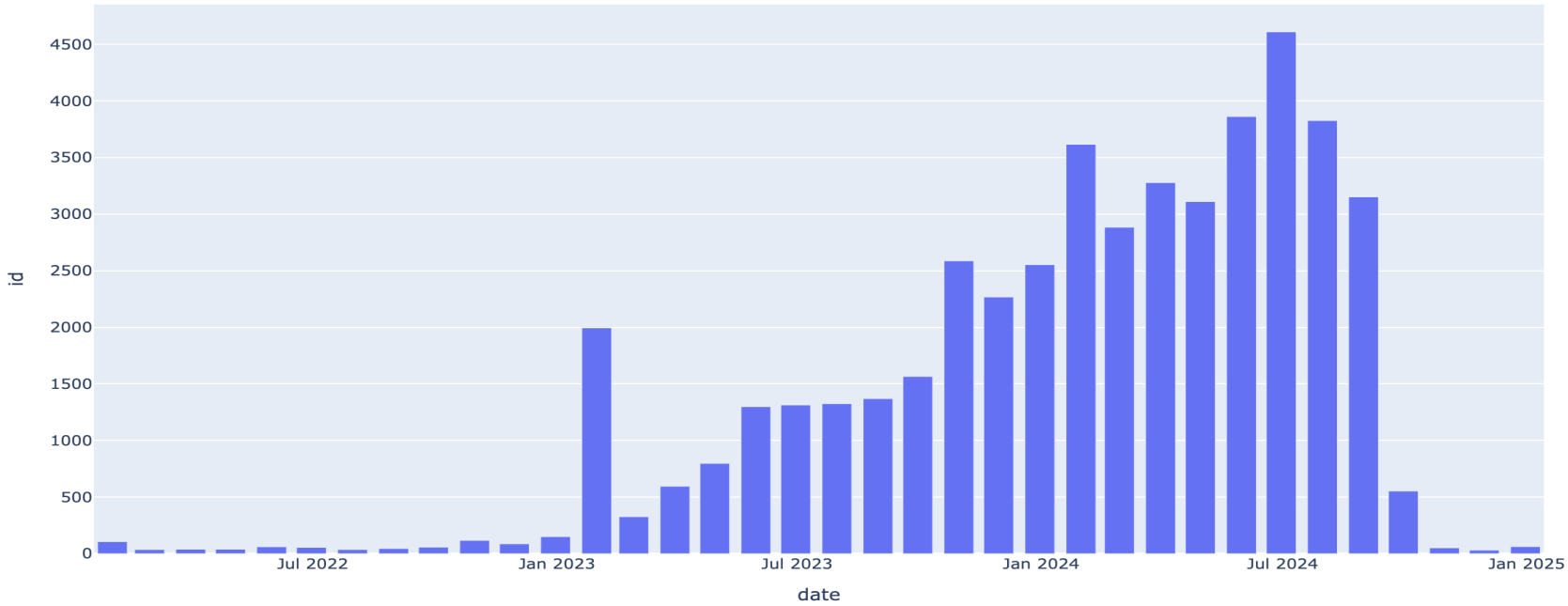


Top 10 Most Cited Publications

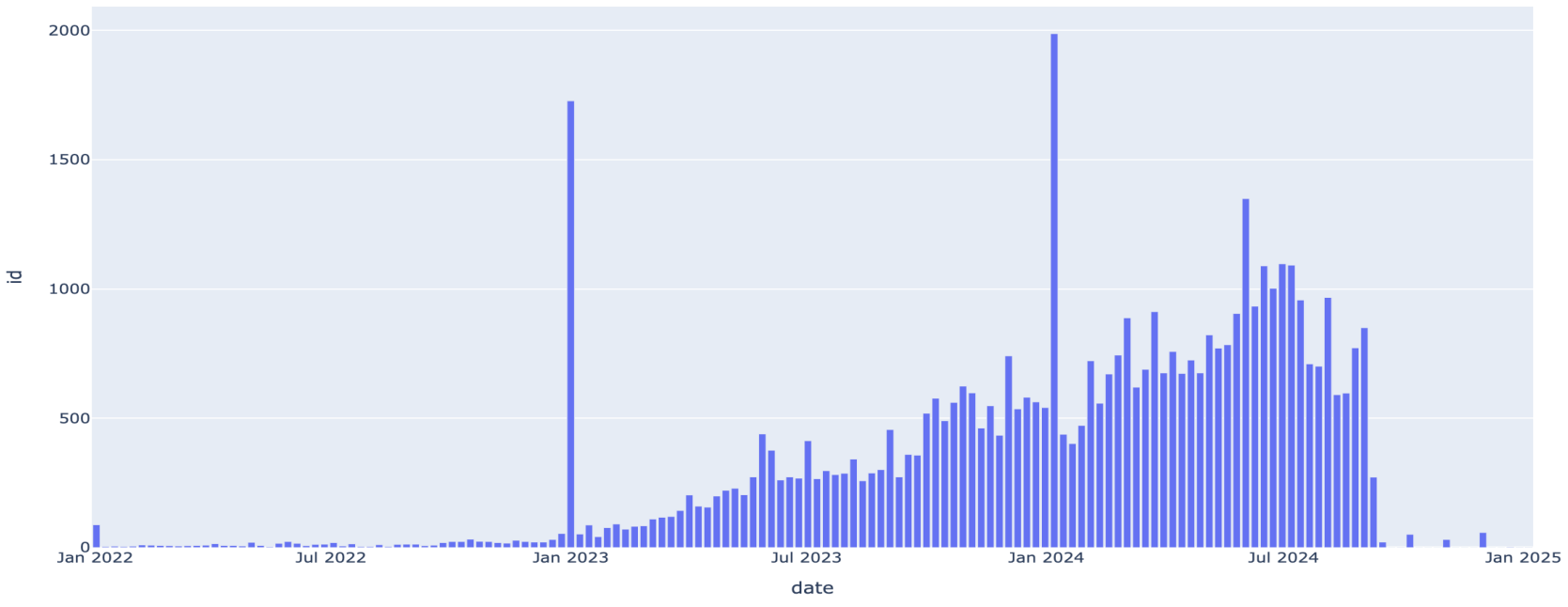
Top 10 Most Cited Publications

	title	doi	times_cited	journal.title	open_access	proportions_female	summed_funding_usd
235	Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models	10.1371/journal.pdig.0000198	1735	PLOS Digital Health	[oa_all, gold]	1	0
6204	“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy	10.1016/j.ijinfomgt.2023.102642	1341	International Journal of Information Management	[oa_all, hybrid]	0	0
303	Evolutionary-scale prediction of atomic-level protein structure with a language model	10.1126/science.ade2574	1318	Science	[oa_all, green]	0	0
9734	ChatGPT for good? On opportunities and challenges of large language models for education	10.1016/j.lindif.2023.102274	1307	Learning and Individual Differences	[oa_all, bronze]	0	0
386	ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns	10.3390/healthcare11060887	1236	Healthcare	[oa_all, gold]	0	0
314	How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment	10.2196/45312	1041	JMIR Medical Education	[oa_all, gold]	0	293812672
1831	ChatGPT: five priorities for research	10.1038/d41586-023-00288-7	880	Nature	[closed]	0	0
2626	ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope	10.1016/j.iotcps.2023.04.003	878	Internet of Things and Cyber-Physical Systems	[oa_all, gold]	0	0
2026	Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum	10.1001/jamainternmed.2023.1838	834	JAMA Internal Medicine	[oa_all, green]	0	523557
209	Large language models encode clinical knowledge	10.1038/s41586-023-06291-2	807	Nature	[oa_all, hybrid]	0	0

Publications - by month

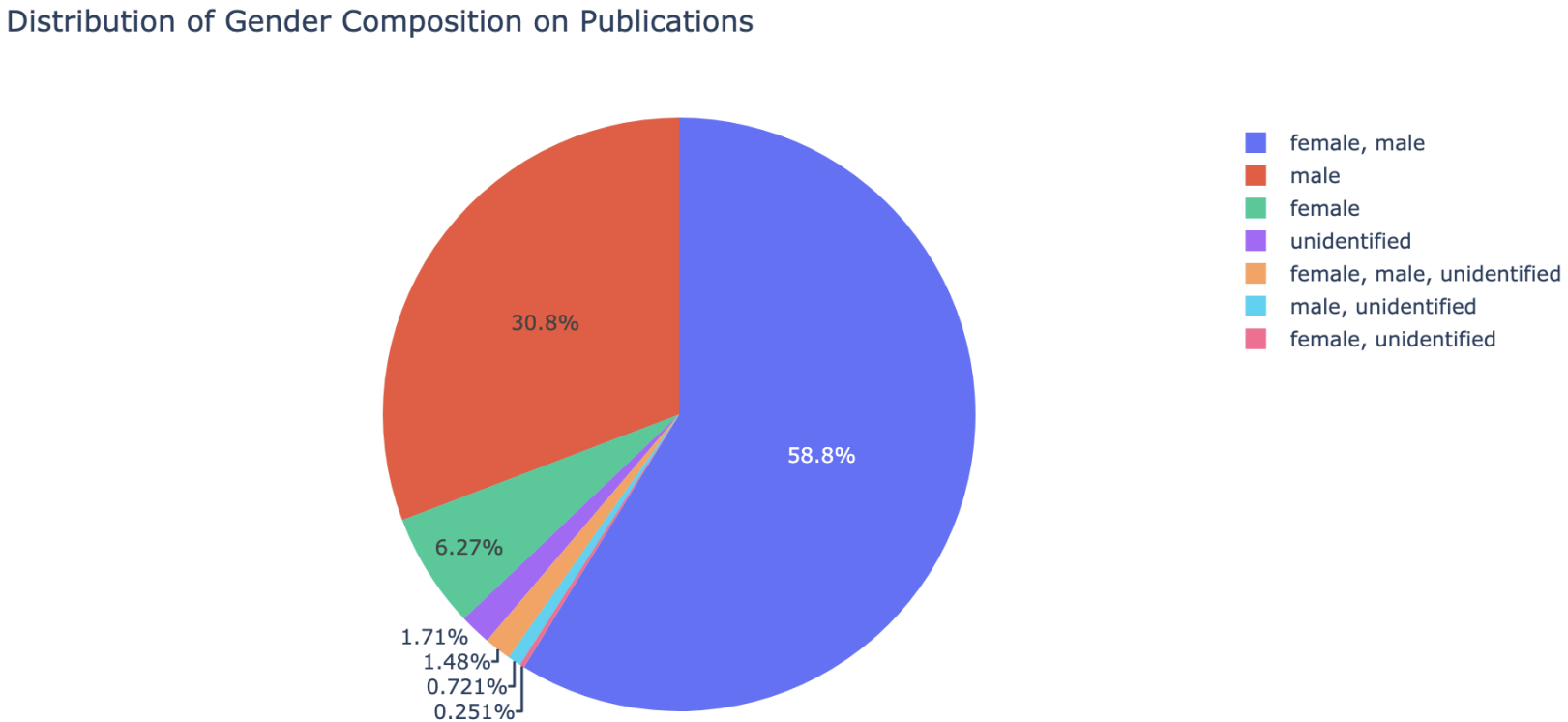


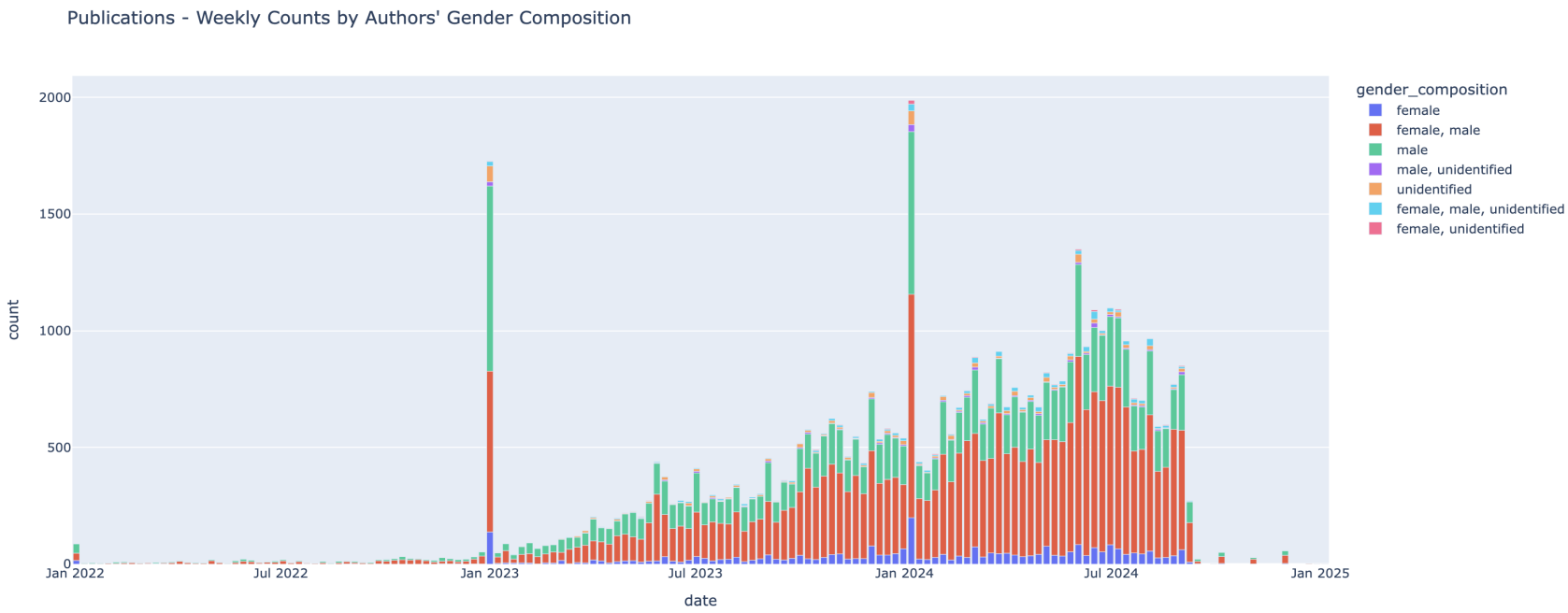
Publications - by week

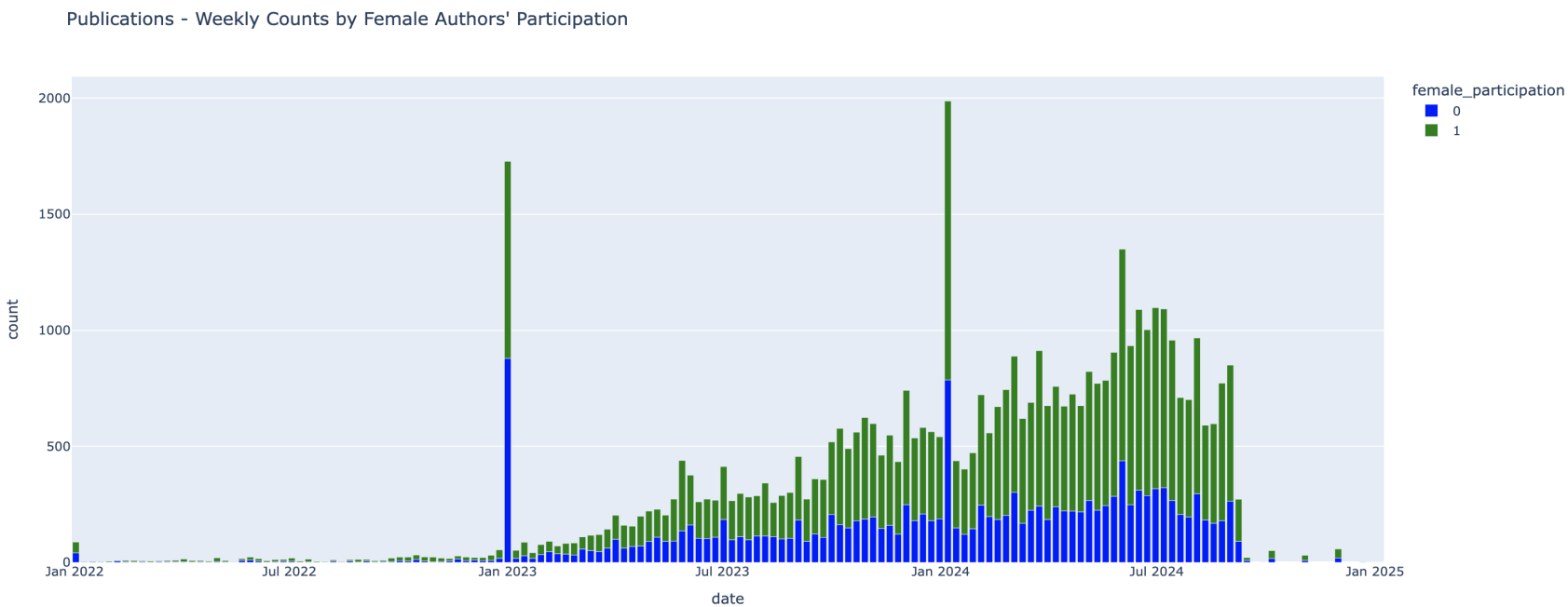


Gender of Authors

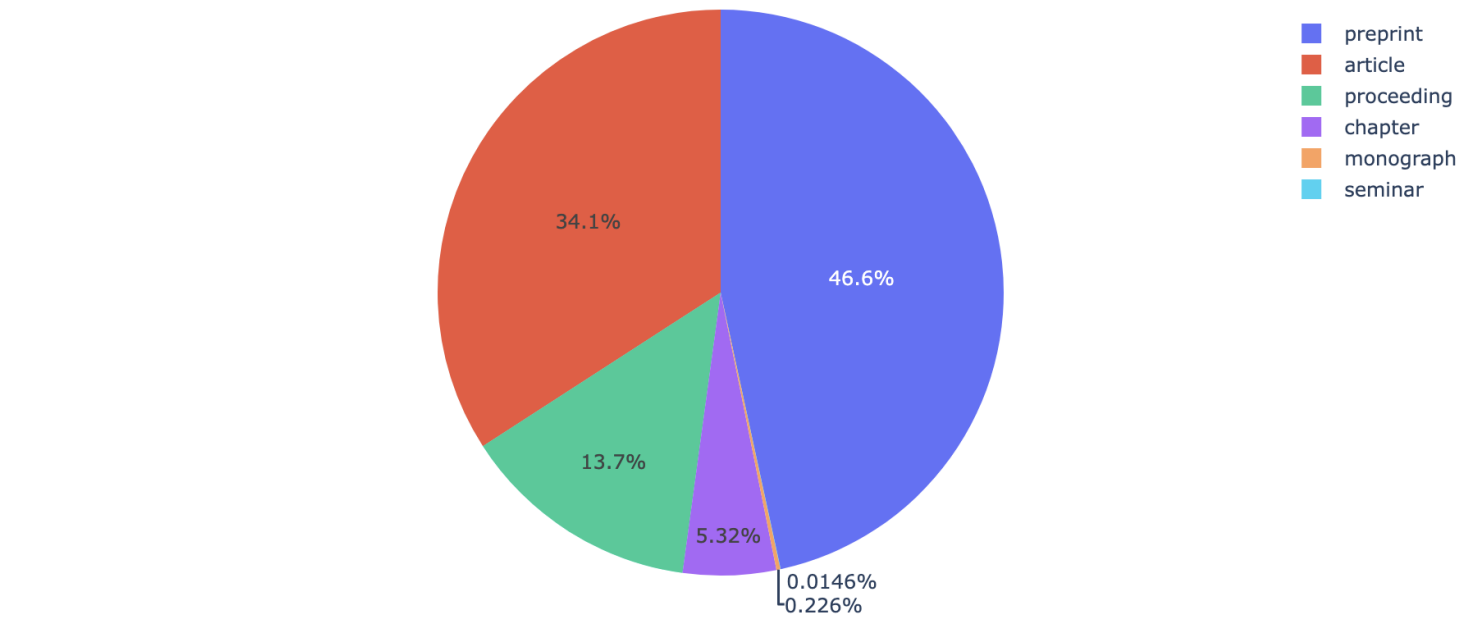
To identify authors' gender, we used **Namsor**, an algorithmic model for the classification of names that contains a repository of 7.5 billion names, including those from 142 ethnicities, 249 countries, and 22 alphabets (`https://namsor.app/about-us/`). Namsor's model recognizes morphemes—the smallest units of construction within languages that help comprise words—to incorporate patterns in naming conventions when assigning a name's gender, ethnic origin, and other elements offered through their service. The accuracy of Namsor's model has been verified by multiple studies and audits, including a 2018 Science-Metrix publication that found it correctly classified the gender of Olympic medalists' names from 25 countries to within 98-99% accuracy (`https://namsor.app/files_to_download_p/science-metrix_bibliometric_indicators_womens_contribution_to_science_report.pdf`).



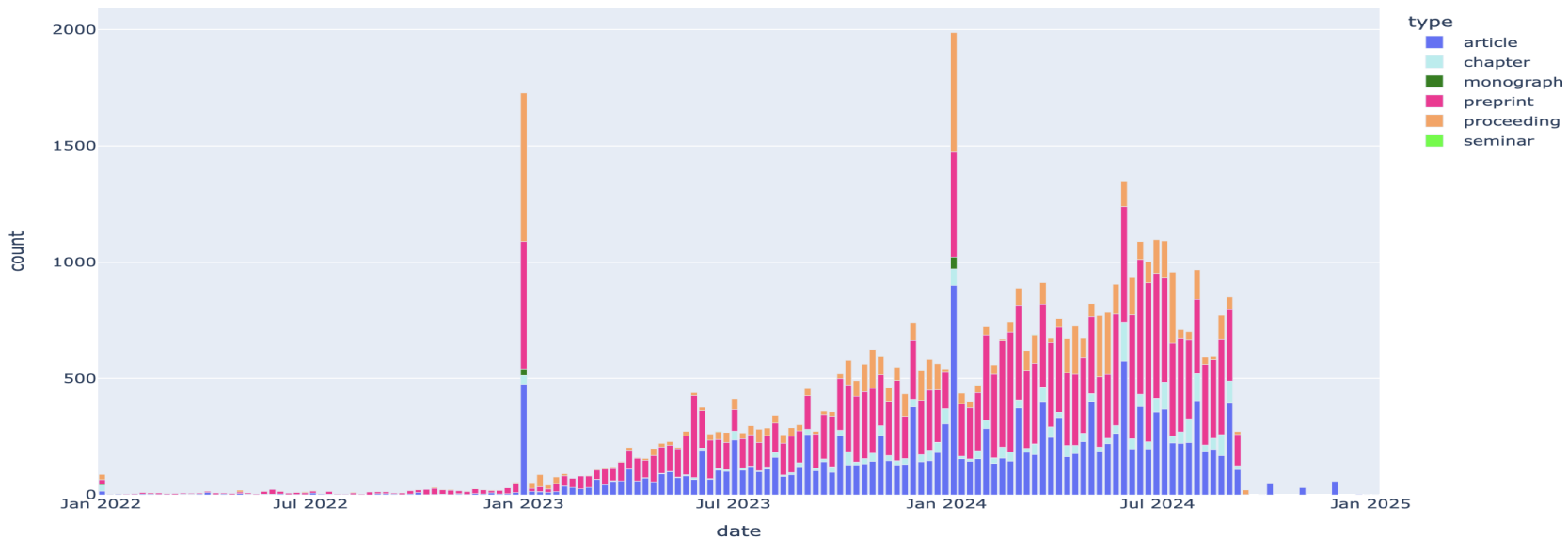

Publications - Weekly Counts by Authors' Gender Composition
gender_composition
female
female, male
male
male, unidentified
unidentified
female, male, unidentified
female, unidentified
count
2000
1500
1000
500
0
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024
Jul 2024
Jan 2025
date


Publications - Weekly Counts by Female Authors' Participation
female_participation
0
1
count
2000
1500
1000
500
0
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024
Jul 2024
Jan 2025
date

Distribution of Publication Types

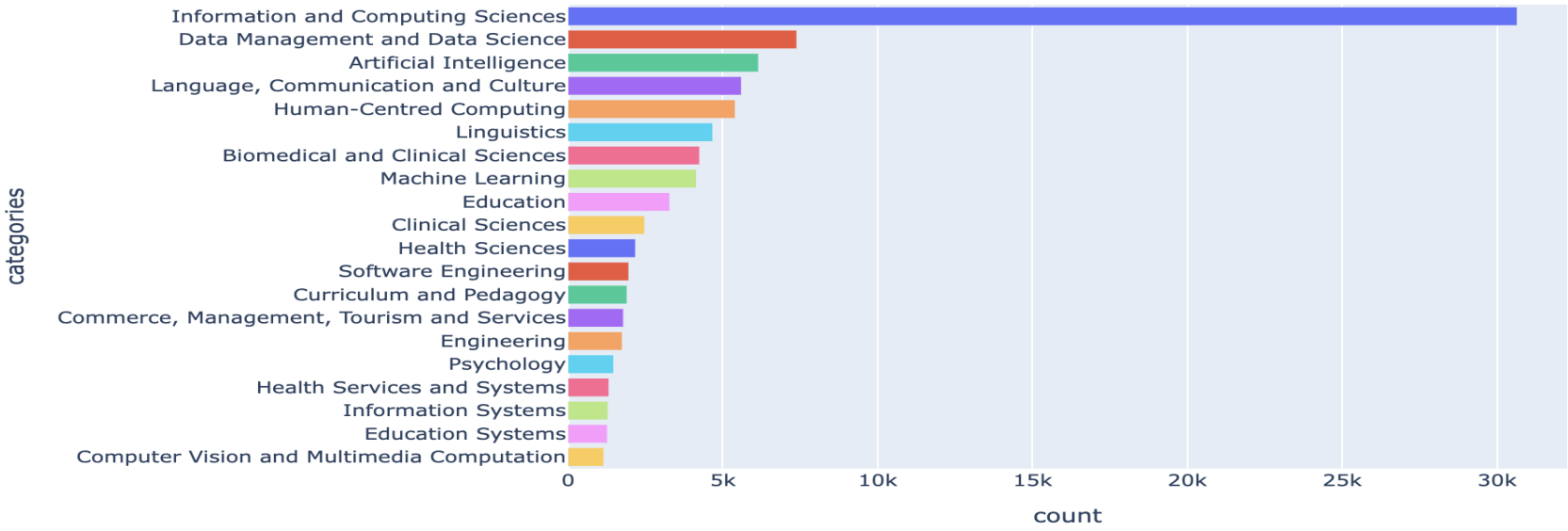


Publications - Weekly Counts by Type

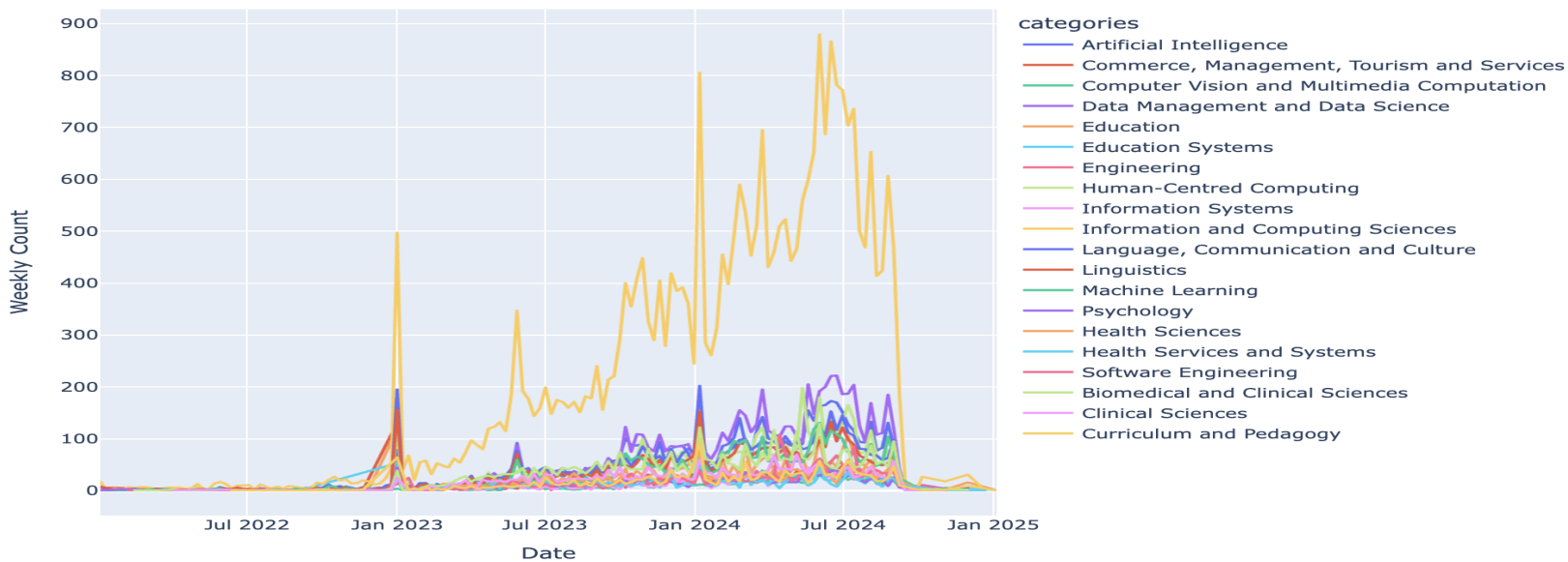


The Dimensions field `category_for` pertains to a classification of the Field of Research (FoR) of publications, which is a classification that aligns with the Australian and New Zealand Standard Research Classification (ANZSRC), which arranges research outputs into a hierarchical structure, where major fields are subdivided into more specific minor fields (`https://plus.dimensions.ai/support/solutions/articles/23000018826-what-is-the-background-behind-the-fields-of-research-for-\protect\penalty\z@classification-system-`).

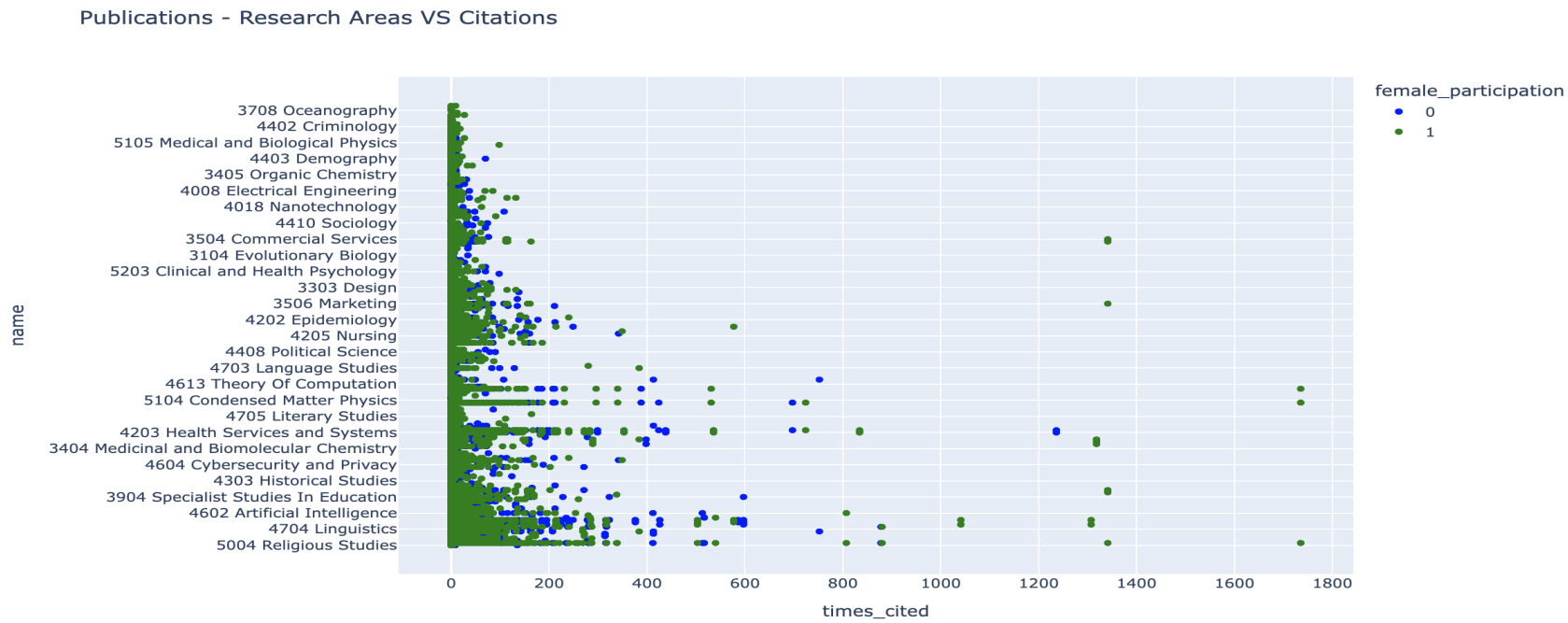
Top 20 Research Areas by Publication Count

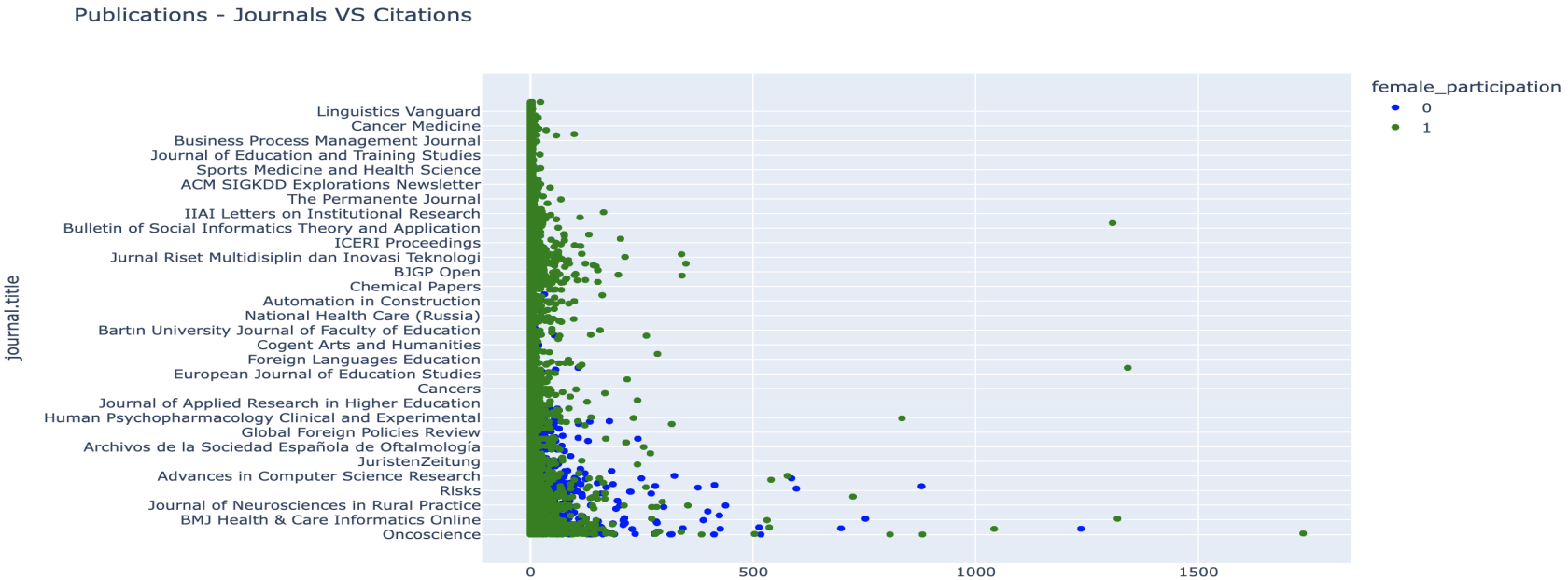


Weekly Evolution of Top 20 Research Areas



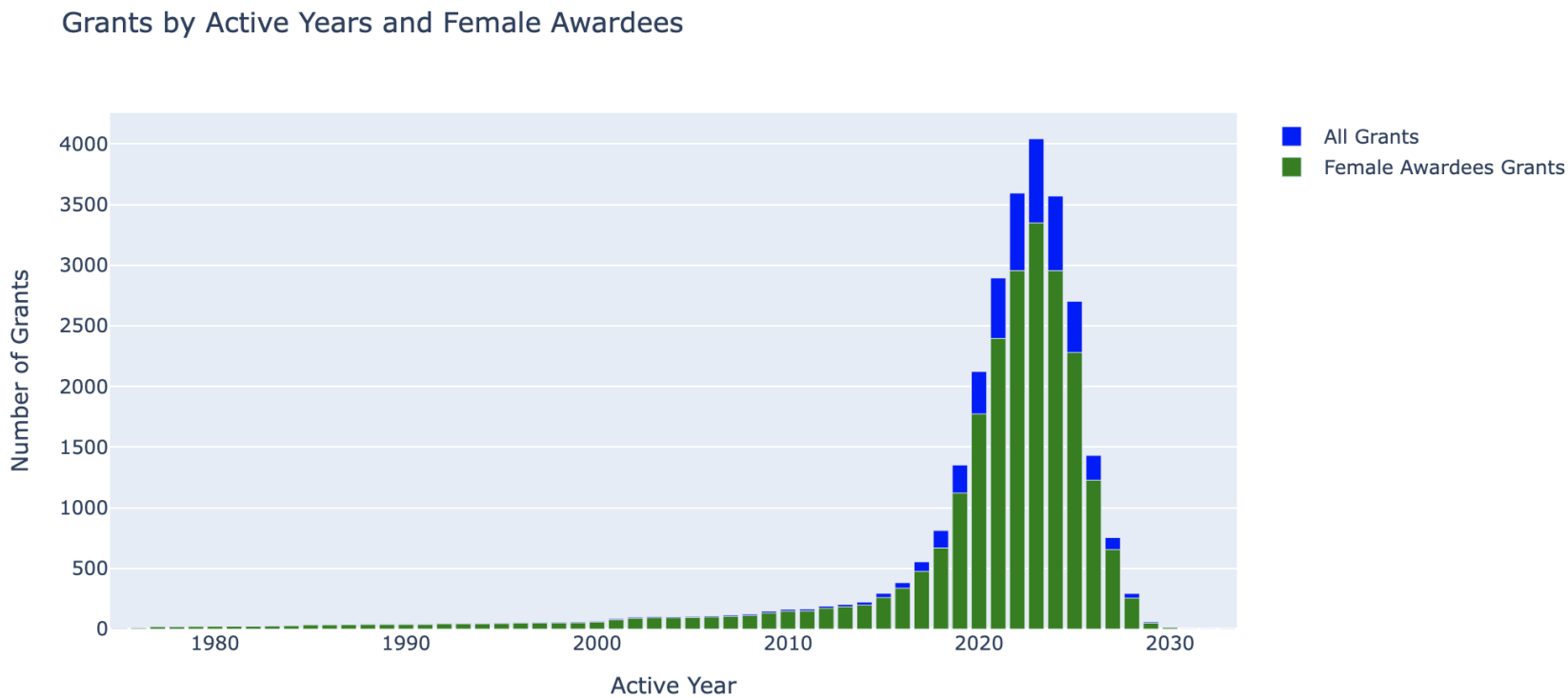
Research Areas and Jounals vs. Citations by Gender



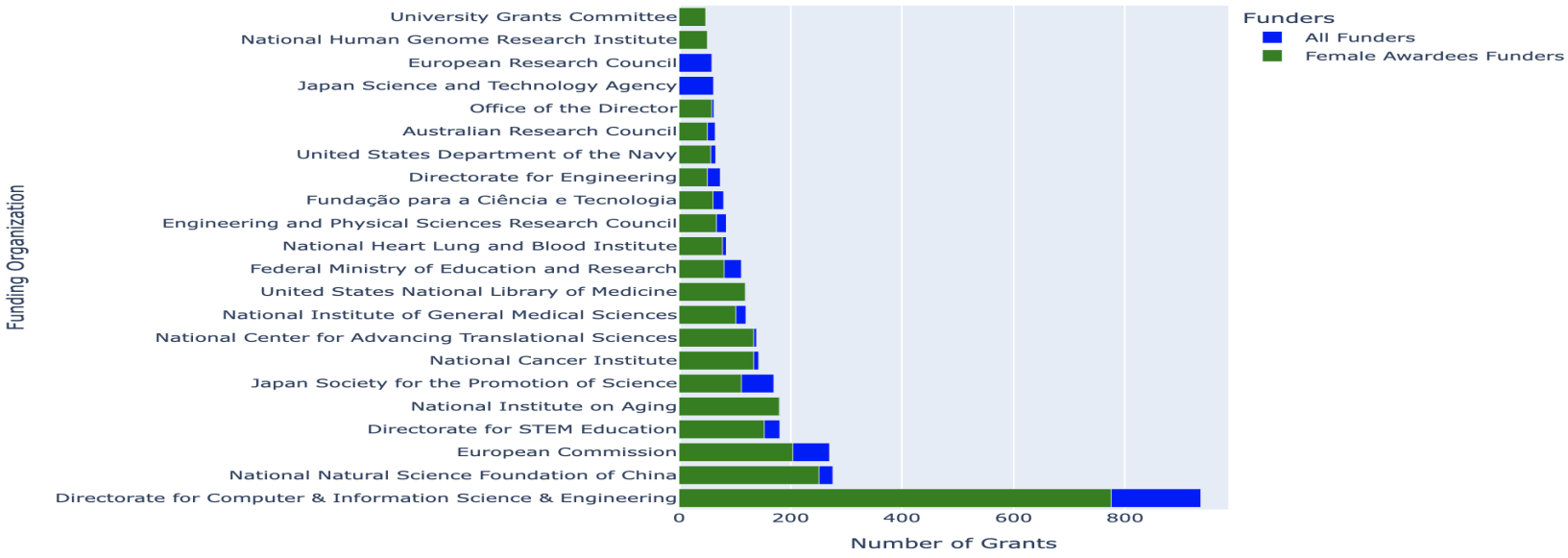


Publications Grants Funding by Gender

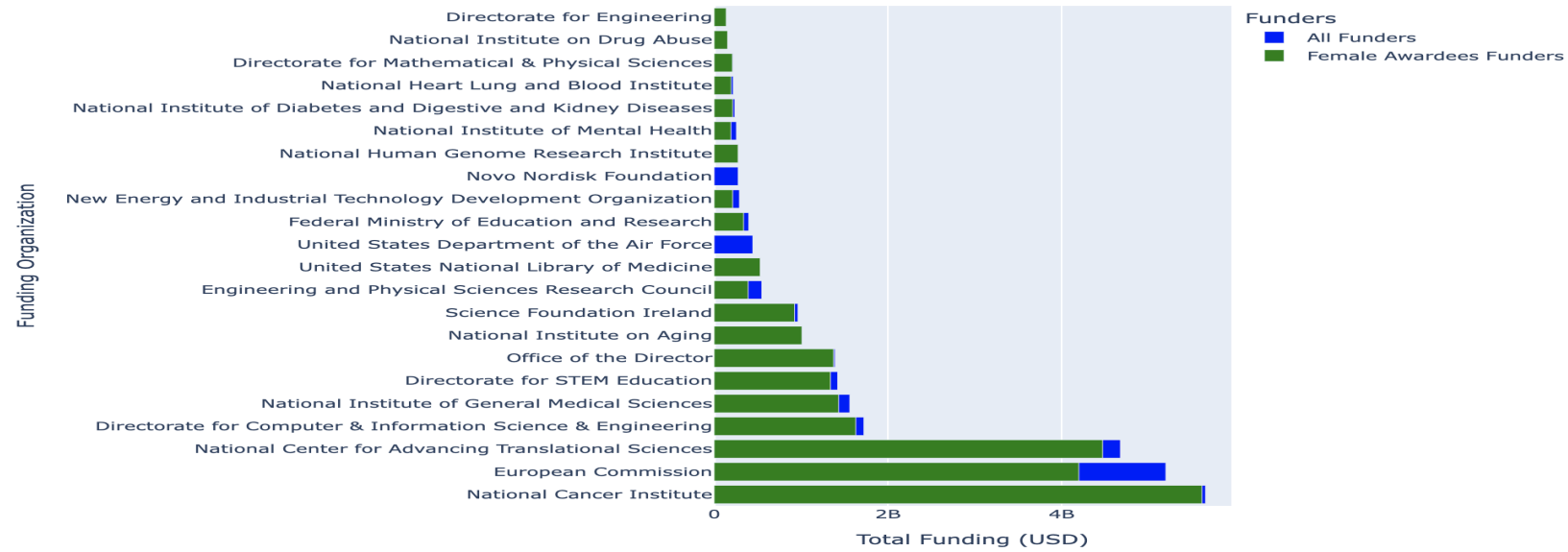
Dimensions.ai maintains a record of grants (in USD) awareded to publication authors (`https://docs.dimensions.ai/dsl/2.0.0/datasource-grants.html`).



Funders and Female Awardees Funders



Funders and Female Awardees Funders by Funding Amount

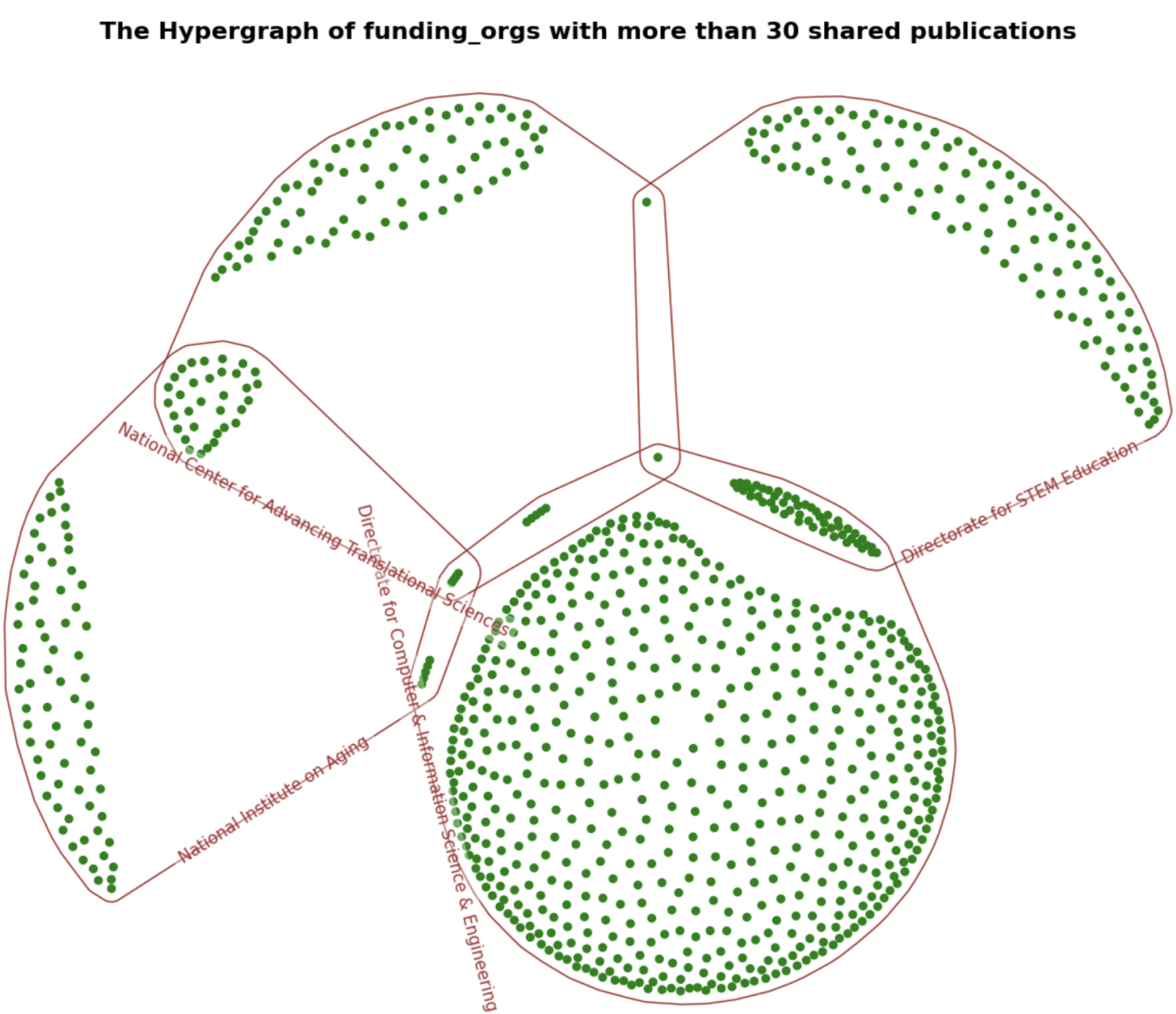


Records of the Top 10 Funders

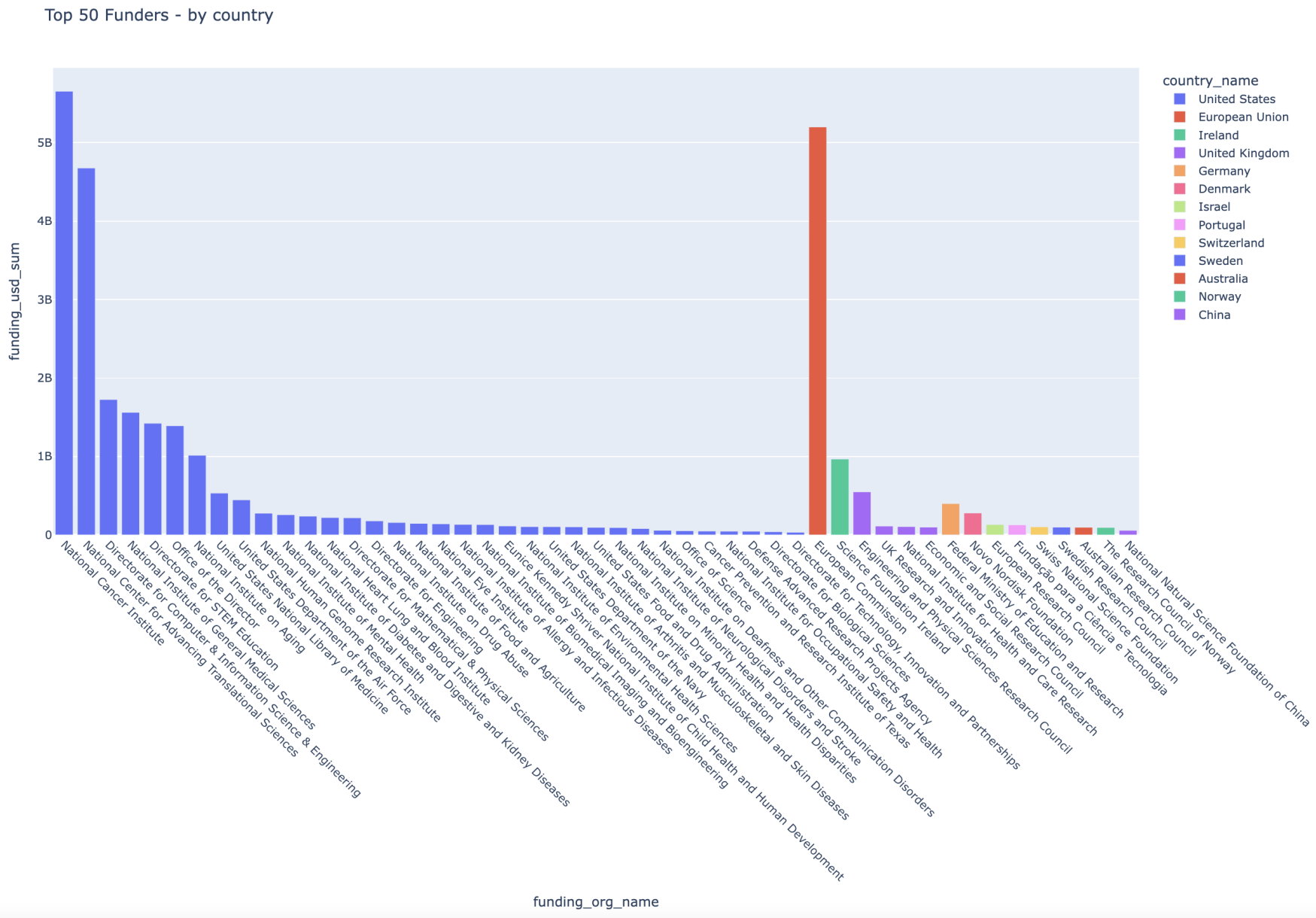
Top 10 Funders

	funding_org_name	funding_usd_sum	count	title	active_years	country_name
69	National Cancer Institute	5,654,672,453	143	The benefits and harms of lung cancer screening in Florida	[2020, 2021, 2022, 2023, 2024]	United States
37	European Commission	5,199,497,391	270	MAchinE Learning for Scalable meTeoROlogy and cliMate	[2021, 2022, 2023, 2024]	European Union
70	National Center for Advancing Translational Sciences	4,675,757,559	139	ENACT: Translating Health Informatics Tools to Research and Clinical Decision Making	[2022, 2023, 2024, 2025, 2026, 2027]	United States
24	Directorate for Computer & Information Science & Engineering	1,724,303,822	936	CAREER: Socially-Aware Language Technologies To Support People in Supporting Others for Better Online Communities	[2022, 2023, 2024, 2025, 2026, 2027]	United States
90	National Institute of General Medical Sciences	1,560,902,303	120	Discovery-Driven Mathematics and Artificial Intelligence for Biosciences and Drug Discovery	[2023, 2024, 2025, 2026, 2027, 2028]	United States
28	Directorate for STEM Education	1,421,547,329	181	AI Institute for Engaged Learning	[2021, 2022, 2023, 2024, 2025, 2026]	United States
113	Office of the Director	1,390,595,659	63	COVID and Translational Science supercomputer (CATS)	[2021, 2022]	United States
94	National Institute on Aging	1,013,681,889	181	Assessing chronic pain using brain entropy mapping	[2022, 2023, 2024]	United States
118	Science Foundation Ireland	965,385,858	28	SFI Centre for Research Training in Machine Learning	[2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026]	Ireland
34	Engineering and Physical Sciences Research Council	548,142,753	85	Maths Research Associates 2021 Oxford	[2021, 2022, 2023, 2024]	United Kingdom

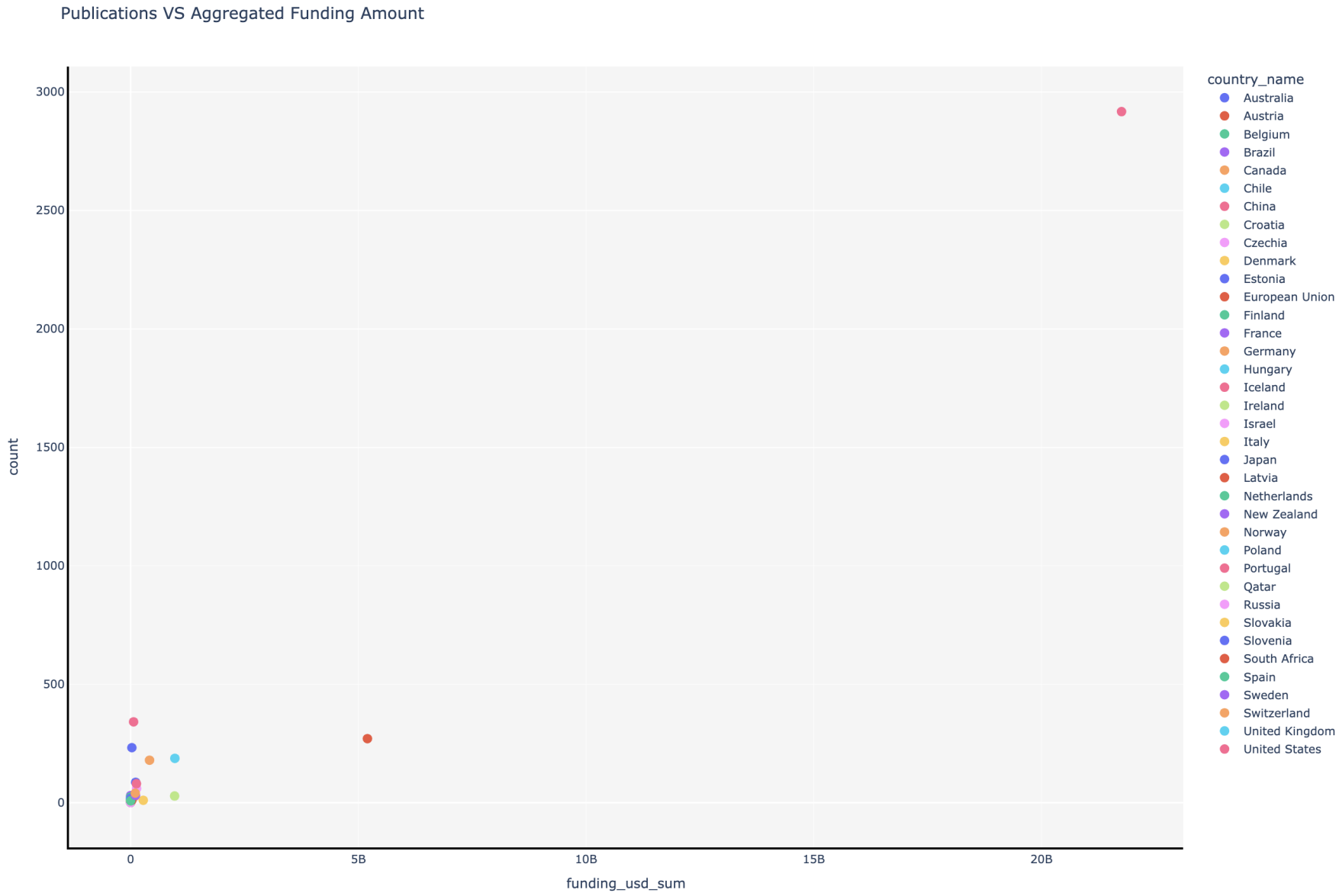
4 Funders Co-Supporting Publications



Countries of the Top 50 Funders



Countries in No. of Publications vs. Aggregated Funding



The Citation Graph

A "citation graph" (or "citation network") is a directed graph where vertices represent publications and edges denote citation relationships between them. Specifically, if publication v cites publication u, an edge is drawn from vertex u to vertex v in the citation graph. More precisely, this defines a "publications citation graph" (i.e., a citation network among publications). Alternatively, one could define an "authors' citation graph," whereevertices represent authors and the edges represent citation relationships between them.

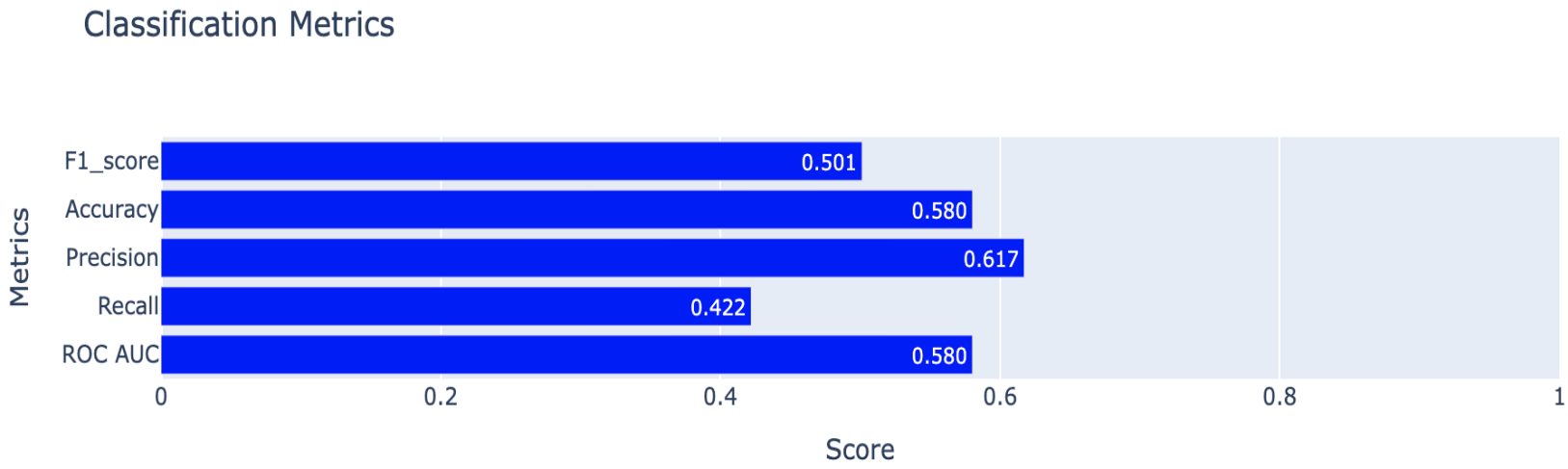
Table: Graph Characteristics

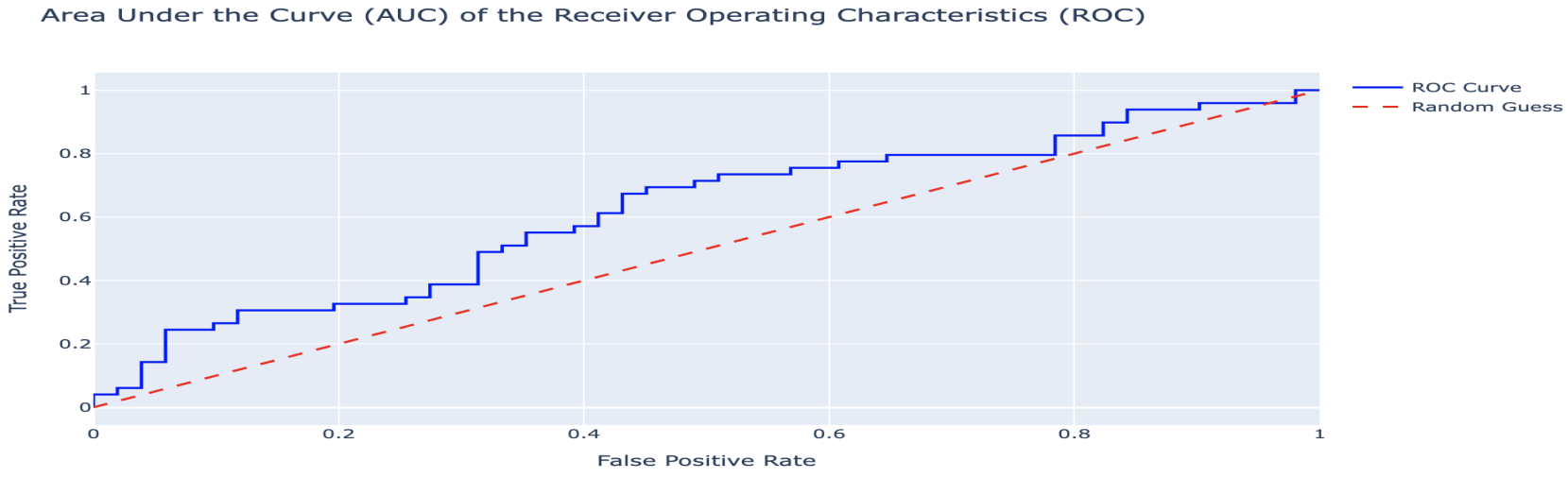
Characteristic	**Value**
Order of the graph	21,426
Size of the graph	107,151
Density	0.0002
Average degree	5.001
Transitivity	0.0039
Reciprocity	0.0019
Number of weakly connected components	205
Order of the largest weakly connected component	20,972
Size of the largest weakly connected component	106,897
Number of strongly connected components	21,331
Order of the largest strongly connected component	12
Size of the largest strongly connected component	60

The Citation Graph as a Directed Hypergraph

- A directed hypergraph is a generalization of a hypergraph in which hyperedges can connect multiple source nodes to multiple target nodes, as opposed to merely containing a set of nodes, which is typical in undirected graphs. We have the following two formal definitions:
  - A **hypergraph** H is a pair of (V, E), where V is a non empty set of nodes and E is a set of hyperlinks (or hyperedges), in which every hyperlink is defined as a set of nodes.
  - A **directed hypergraph** H is a pair of (V, E), where V is a non empty set of nodes and E is a set of (directed) hyperlinks (or directed hyperedges), in which every (directed) hyperlink is an ordered pair (S, T) of two sets of nodes, where S is a set of source nodes and T is a set of target nodes.
- AN IMPORTANT PROBLEM: **Hyperlink prediction in citation graphs considered as directed hypergraphs**.

Hyperlink Prediction in the Citation Directed Hypergraph via the CHESHIRE method





Here, the CHESHIRE method yields unsatisfactory results; therefore, alternative methods should be explored.