Metabolomics: Mapping Evolution and Trends in the Field with AI Tools

Georgia Institute of Technology

Fernández Lab @ GT

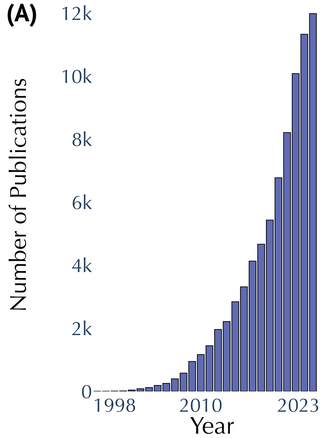
Olatomiwa O. Bifarin[1], Varun S. Yelluru[2], Aditya Simhadri[2], and Facundo M. Fernández[1,3]

[1] School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia, United States
[2] School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, United States
[3] Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, Georgia, United States

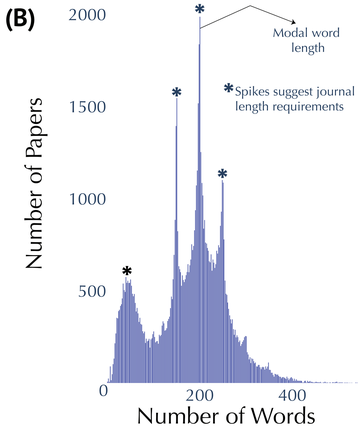
Poster Tour

Publication Analytics

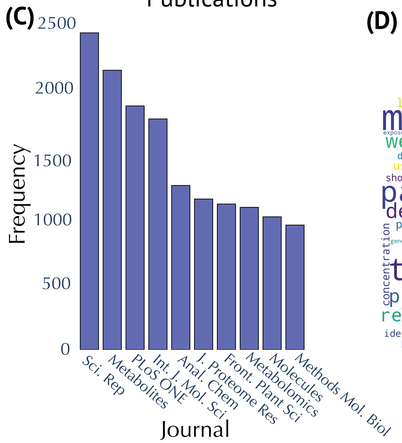
Annual *Publication* Growth



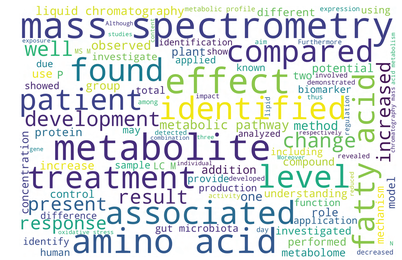
Abstract *Length* Distribution



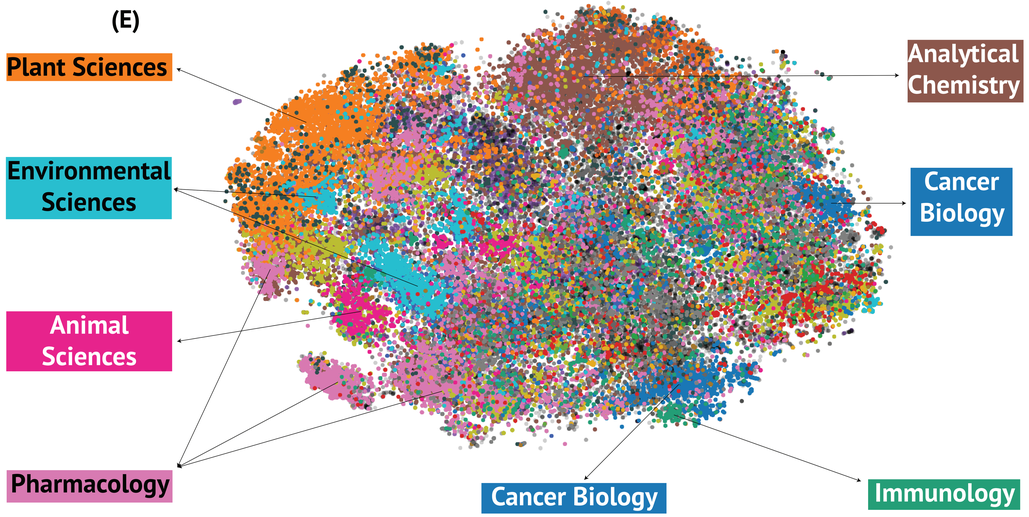
Top Journals by Metabolomics Publications



Frequently Used Words in Abstract



2D tSNE Embeddings of Metabolomics Papers



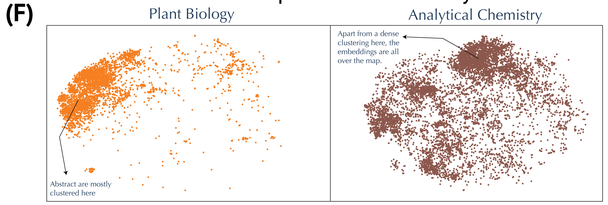
In Brief

Thousands of metabolomics papers are published annually, making it challenging to stay updated. To address this, we hypothesized that unsupervised word embeddings could capture latent knowledge from metabolomics literature. Using PubMed abstracts containing "metabolomics" or "metabonomics," we compiled a dataset of over 82,000 entries from 1998 to January 2024. The number of publications has increased yearly since the late 1990s **(A)**. We conducted publication analytics and found the modal abstract length to be 200 words, with spikes suggesting journal length requirements **(B)**. The top journals for metabolomics publications are presented in **(C)**, with Scientific Reports leading, followed by Metabolites and PLoS ONE. Frequently used words in abstracts is shown in **(D)** highlighting mass spectrometry. For embedding generation, abstracts were tokenized and processed through the PubMedBERT model to create embeddings, yielding 768-dimensional vectors. These were reduced to two dimensions using tSNE **(E)**, with labels inferred from journal titles. Analysis revealed distinct clusters for plant sciences and dispersed clusters for analytical chemistry **(F)**. Overlapping fields include toxicology and environmental sciences **(G)**, cancer biology and immunology (not shown)

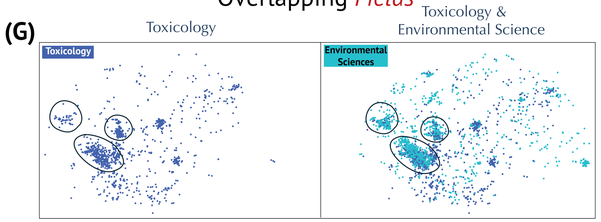
The methodology also enables keyword searches to uncover trends, such as COVID-19 metabolomics publications clustering **(H)**. Investigation of embeddings by publication year showed no apparent trends **(I)**. Future work will explore the evolution of research topics and techniques in metabolomics using these embeddings. Furthermore, with almost half of the embeddings unlabeled (n = 41,721), future directions include topic modeling with open-source models like LLaMa3.

Trends in Embeddings

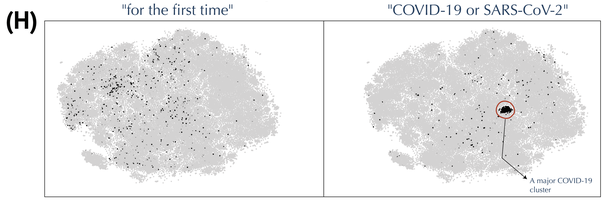
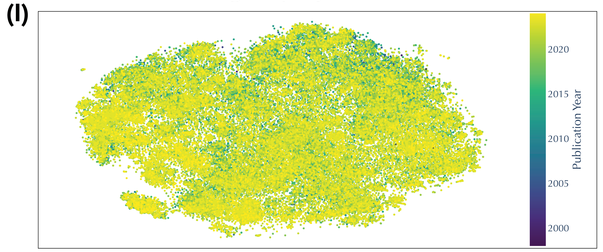
Isolated and Dispersed Clusters by *Fields*



Overlapping *Fields*



Keywords *Trends*



Embedding Map by *Publication* Year