Yale SCHOOL OF MEDICINE & Yale SCHOOL OF PUBLIC HEALTH

Literature Visualization of the Shifting PFAS Exposure Research Landscape

Weipeng Zhou, PhD[1], Vipina K. Keloth, PhD[1], Rixin Wang, PhD[1], George Michel, MS[1], Haseena Rajeevan, PhD[1], Casper Zhang, BS[1], Eric Gossett, MS[1], Hua Xu, PhD[1], Nicole Deziel, PhD[2], Kei-Hoi Cheung, PhD[1,2]
[1]Yale School of Medicine; [2]Yale School of Public Health, New Haven, CT, USA

Introduction

- **Per- and polyfluoroalkyl substances (PFAS)** are a large, chemically diverse group of synthetic compounds that have garnered global attention due to their persistence, widespread human and environmental exposure, and potential toxicity[1].
- They have been used in numerous industrial and consumer products such as **food packaging, textiles, coatings, and firefighting foams**.
- PFAS are often referred to as **"forever chemicals"** because their unique chemical structure leads to their persistence in the environment and the human body.
- PFAS exposure is **nearly ubiquitous**, with evidence linking certain compounds to metabolic, developmental, and immune system effects.
- **Definitional and regulatory complexity** adds to the challenge:
  - The Organization for Economic Co-operation and Development or **OECD (2021)** broadened the PFAS definition to **>7 million compounds**, while other inventories list only thousands.
  - **Recent policy changes**, including shifts in **EPA drinking-water regulation** and federal funding priorities, are reshaping the PFAS research landscape.
- These dynamics call for the need to **track evolving research themes** and identify **emerging or neglected compounds**.

(a)

Figure 1. Each dot in (a) corresponds to a PubMed article. The colors represent different time periods. The articles between 2010 and 2025 are clustered into different themes including "water pollution". Articles belonging to a particular theme (e.g., pregnancy outcome) can be selected for a natural language query as shown in (b). The query returns top 5 PFAS chemicals mentioned in the selected articles and these PFAS chemicals are also included in EPA's OECD PFAS list. They are displayed as dots in different colors as shown in (c).

Objective

To build a **data-driven, interactive visualization ("MedViz")** that:

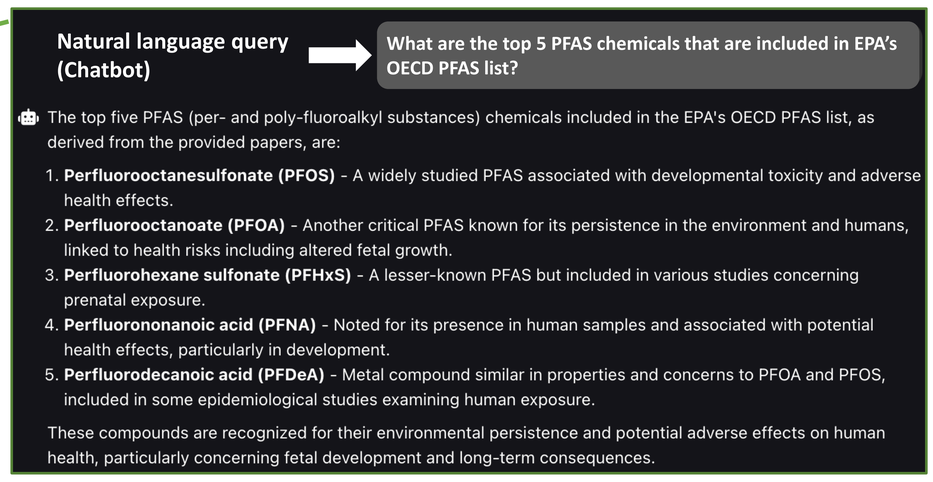
- Maps **five decades (1974–2025)** of PFAS-related biomedical literature.
- Reveals **research trends, clusters, and gaps** in the scientific understanding of PFAS.
- Enables stakeholders to **explore and reorient research priorities** in response to regulatory or funding shifts.

Methods

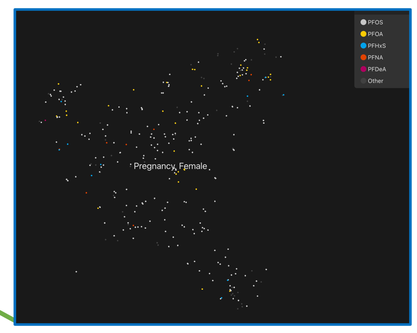
- **Data Source:** PubMed abstracts from **1974–2025**.
- **Filtering:** Applied a curated set of **PFAS-specific keywords²** to titles and abstracts → retained **16,597 relevant articles**.
- **Text Representation:**
  - Concatenated title + abstract for each paper.
  - Generated **384-dimensional text embeddings** using **bge-small-en-v1.5** model.
- **Dimensionality Reduction:**
  - Applied **t-SNE** to project embeddings into **2D space** for visualization in *MedViz*.
- **Semantic Clustering:**
  - Conducted **hierarchical clustering** on article **MeSH terms** to group related research topics.
  - Clusters aligned with PFAS research domains (e.g., *toxicology*, *environmental monitoring*, *analytical chemistry*, *epidemiology*).
- **Interactive Platform:**
  - Visualized using **MedViz** (https://medviz.org).
  - Users can query, zoom, and filter by **year, topic, or PFAS compound**, supporting exploratory analyses.

Results

- **Comprehensive corpus:** 16,597 PFAS-related studies spanning 50 years.
- **Thematic clusters** correspond to key subfields:
  - *Analytical detection and quantification*
  - *Toxicity and bioaccumulation*
  - *Epidemiological and exposure studies*
  - *Environmental distribution and remediation strategies*
  - *Emerging PFAS and novel fluorinated compounds*
- **Temporal patterns:**
  - Pre-2000 research dominated by **analytical chemistry**.
  - Post-2010 surge in **health and exposure studies** following growing public health concern.
  - **Recent years (2020–2025)** show increasing focus on **regulation, risk communication, and data integration.**
- The **interactive visualization** highlights under-explored chemical subclasses and research areas with limited coverage despite high exposure potential.



(b)



(c)

SCAN ME

For Best Compatibility
Please Open with Chrome

Discussion

- **Policy shifts and definitional updates** substantially reshape research attention.
- *MedViz* enables dynamic tracking of these changes by providing an **up-to-date, visual knowledge map**.
- Embedding-based clustering uncovers **latent relationships** between studies that span multiple disciplines.
- The combination of **semantic embedding + ontology-based clustering** bridges gaps between environmental science, toxicology, and biomedical research communities.
- This approach demonstrates the power of **AI-driven literature synthesis** to keep pace with evolving regulatory landscapes.

Conclusions & Future Work

MedViz supports researchers and policymakers in:

- Identifying **critical knowledge gaps** and **understudied PFAS compounds**.
- Aligning research investment with **public health relevance**.
- Tracking how **funding and policy shifts** shape the direction of PFAS science.

Future directions:

- Integrating **full-text analysis** and **citation networks**.
- Expanding to **other environmental contaminants**.
- Building **temporal dashboards** to monitor ongoing changes in scientific focus.

Acknowledgments

This project is funded in part by NIH grants: R24ES036135 and P42ES033815.

References

1. Cousins IT, DeWitt JC, Juliane Glüge J, et al. (2020) Strategies for grouping per- and polyfluoroalkyl substances (PFAS) to protect human and environmental health, Environ. Sci.: Processes Impacts, 22, 1444-1460.
2. Kobayashi Y, Uchida T, Inoue T, et al. (2024) A Comprehensive Analysis of the per- and poly-fluoroalkyl substances (PFAS) research landscape through AI-assisted text mining, J Hazard Mater. Lett., Vol 5, 100121.