



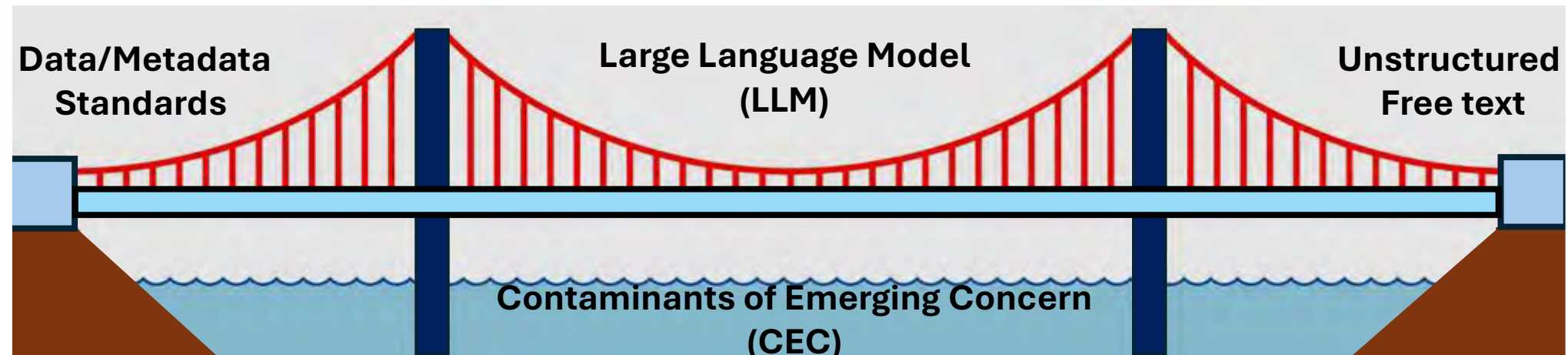
Large Language Model as a Bridge over Unstructured Contaminated Water Data

Kei-Hoi Cheung, PhD

Biomedical Informatics & Data Science (BIDS)

Yale School of Medicine

LLM as a bridge between unstructured data and standards



Automatically extract CEC exposure metadata from literature

Introduction



Contaminants of Emerging Concern (CECs) are natural or man-made chemicals not commonly monitored in the environment



CECs in environmental media (e.g., water) may pose significant public health risks



Systematic reviews of literature can help identify and characterize these risks



The rapidly expanding volume of publications makes manual review increasingly challenging



To address this, a scalable, automated pipeline using Large Language Models (LLMs) was developed to extract metadata from publications related to CECs

Why has the use of LLM in environmental evidence synthesis been limited?



Non-deterministic outputs, hallucinations, and the need for scalable yet reliable evaluation strategies



Need for expert oversight to interpret complex environmental studies



Significant technical knowledge is required



High energy consumption for LLMs presents significant environmental impacts

PECO Framework

- **Population:** The specific group of people or subjects being studied (e.g., adults, workers, pregnant women).
- **Exposure:** The agent or risk factor the population is exposed to (e.g., water/air pollution, chemical).
- **Comparator:** The group or level of exposure to which the target group is compared (e.g., low vs. high exposure).
- **Outcome:** The specific health effects or results being measured (e.g., cancer, respiratory disease, obesity)



Reporting guidelines (checklists)

- A reporting guideline is a simple, structured tool for researchers to use while writing manuscripts reporting research findings
- It provides a minimum list of items (entities) needed to ensure the results can be:
 - understood by a reader
 - replicated by a researcher
 - used by a policy maker to make an informed decision
 - included in a systematic review and possibly meta-analysis



A PECO-aligned checklist

Study
population

Study year

Contaminant

Medium

Source

Location

Biospecimen

Health
outcome

Podder et al. (2026) Green prompt engineering for sustainable generative AI, *Env. Sci. & Ecotechnology*, 30: 100684



Six best practices

- #1: Variational prompting
- #2: Avoid local lingua
- #3: Avoid ambiguities
- #4: User modeling
- #5: Specify length
- #6: Specify format



Results

32–48%

energy and CO₂ reduction

Consistent savings across
proprietary and open-source
large language models

100+ scenarios;
300+ prompts evaluated

No model retraining or
hardware changes required

Use case: PFAS exposure in drinking water



PFAS is a widely studied class of persistent, synthetic compounds that have garnered global attention due to their widespread human exposure and potential toxicity



PFAS have been extensively used in industrial and consumer applications (e.g., food packaging, textiles, and firefighting foams)



PFAS are called “forever chemicals”, as they are characterized by strong carbon-fluorine bonds that confer environmental persistence and bioaccumulation

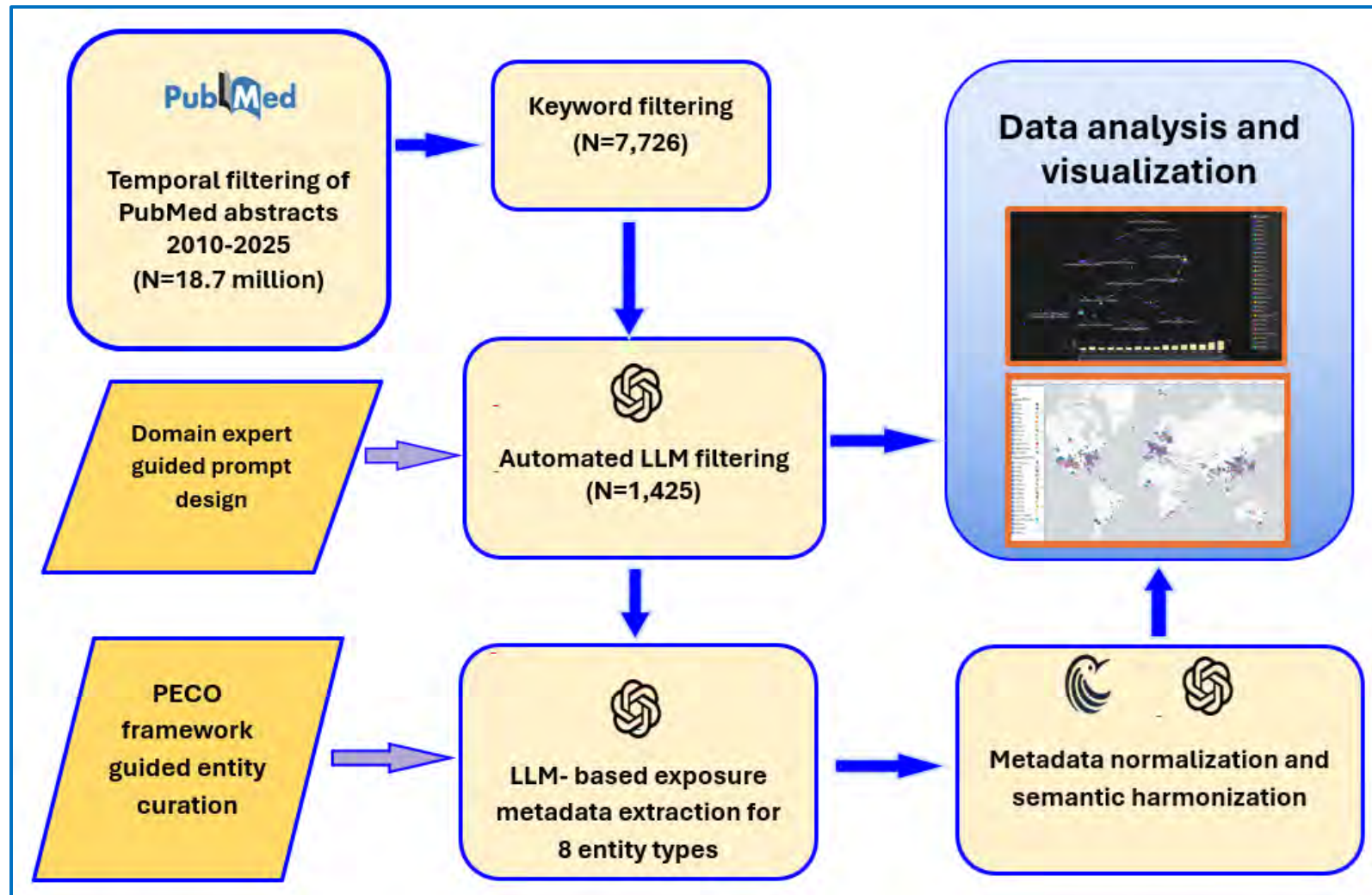


Human exposure to PFAS is nearly ubiquitous and the health effects associated with PFAS exposure is widespread



Among different exposure sources, contaminated drinking water is a major, widespread source of PFAS exposure

Schematic overview of the LLM pipeline



Green prompt example 1 (user modeling)

You are an expert environmental scientist specializing in studying PFAS water contamination events.

Your knowledge should be informed by the following key references on PFAS to help you better understand the context of your task:

1. Sunderland, E.M., et al. (2019). A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. *Journal of Exposure Science & Environmental Epidemiology*.
2. Fenton, S.E., et al. (2021). Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. *Environmental Toxicology and Chemistry*.
3. National Science and Technology Council (NSTC). (2023). Per- and Polyfluoroalkyl Substances (PFAS) Report. Joint Subcommittee on Environment, Innovation, and Public Health.

Your task is to extract entities exactly as they appear in the text — without normalization, inference, or modification.

Read the provided title and abstract carefully.

Green prompt example 2 (specify format)

All fields must be JSON lists of strings as the format below
If there is more than one in your extracted field, you must list them all separated by ',' character (e.g. ["human", "mice"])
If the extracted value is "none", output an empty list [].
The JSON must contain no comments, no trailing text, and no extra explanation.
The JSON object must be the ONLY content in your final answer.

Example Output:

```
{  
  "study_population": ["children"],  
  "study_year": [],  
  "biospecimen": ["kidney", "blood serum"],  
  "outcome": ["functional thyroid disease", "hyperthyroidism",  
  "hypothyroidism"],  
  "location": ["China"],  
  "pfas": ["PFBA", "PFOA"],  
  "medium": ["water"],  
  "source": []  
}
```

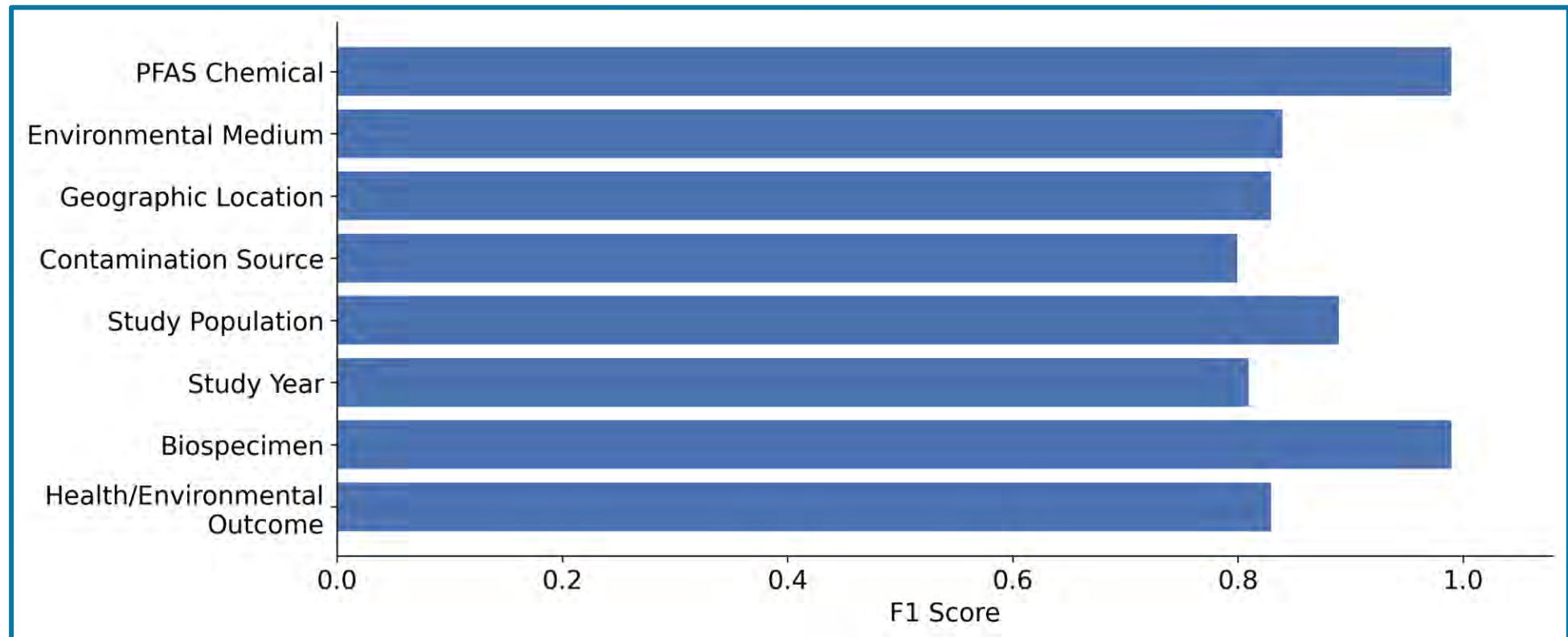
We applied **GPT-5.2-Chat14** to extract PFAS exposure metadata from **1,425 abstracts**

Entities extracted

Entity Type	Description	# Entities Extracted
PFAS Chemical	Specific per- and polyfluoroalkyl substances mentioned (e.g., PFOS, PFOA, PFHxS).	3,032
Environmental Medium	Environmental media where PFAS was detected (e.g., drinking water, groundwater, river water)	2,745
Geographic Location	Geographic locations associated with the study (e.g., Ohio, Danube River, Beijing)	2,456
Contamination Source	Sources of PFAS contamination (e.g., industrial discharge, AFFF, wastewater treatment plant)	1,657
Study Population	Population studied (e.g., residents, children, pregnant women)	307
Study Year	Time period of the study (e.g., 2007–2010, May 2007 until August 2008)	201
Biospecimen	Biological specimens collected (e.g., blood samples, serum, urine)	140
Health/Environmental Outcome	Health or environmental outcomes measured (e.g., serum PFAS concentration, health risk, estimated daily intake)	460
Total		10,998

Evaluation

- The LLM prompting and evaluation was conducted under an iterative framework.
- The LLM performance was evaluated against human annotations on 100 abstracts
- The LLM achieved an average exact match F1 score of 0.92 (macro-averaged across eight entity types)



Metadata normalization and semantic harmonization



The extracted entities are mostly in the form of free text (e.g., bottled water, bottled mineral water, bottled water samples)



Standardizing them can help facilitate downstream analysis



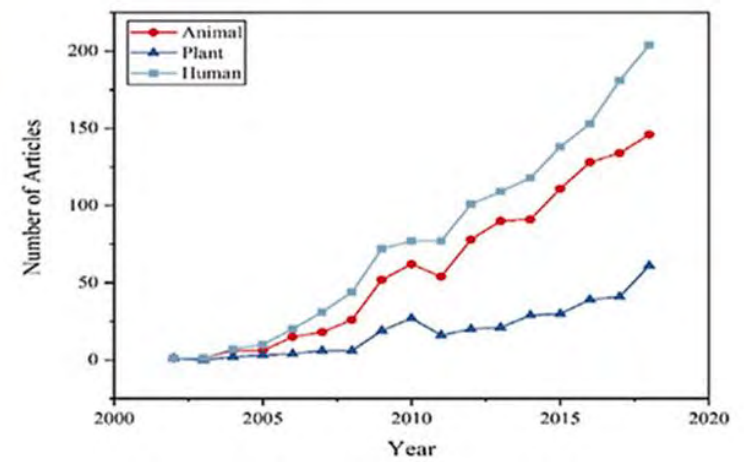
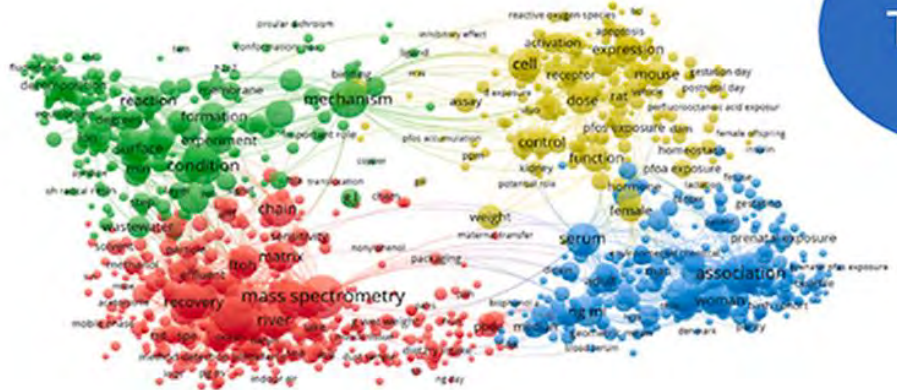
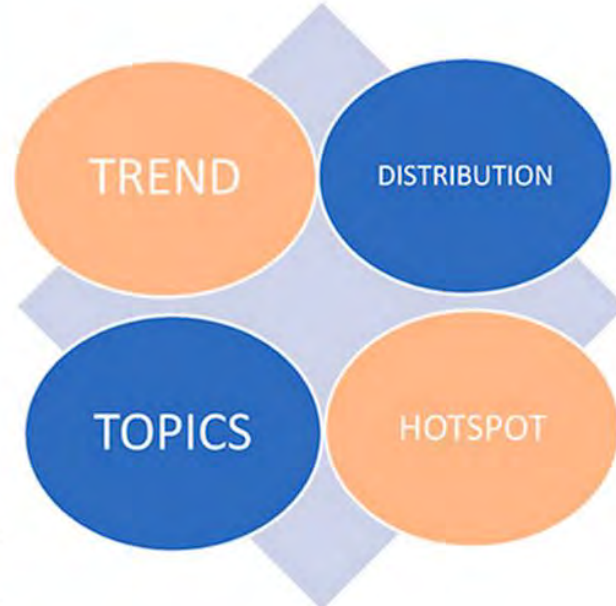
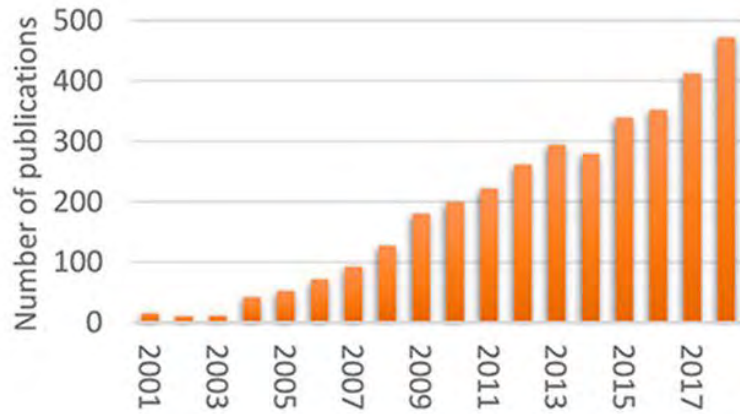
Our initial approach for standardization is to map them to Unified Medical Language System (UMLS) concepts

UMLS mapping across 100 abstracts

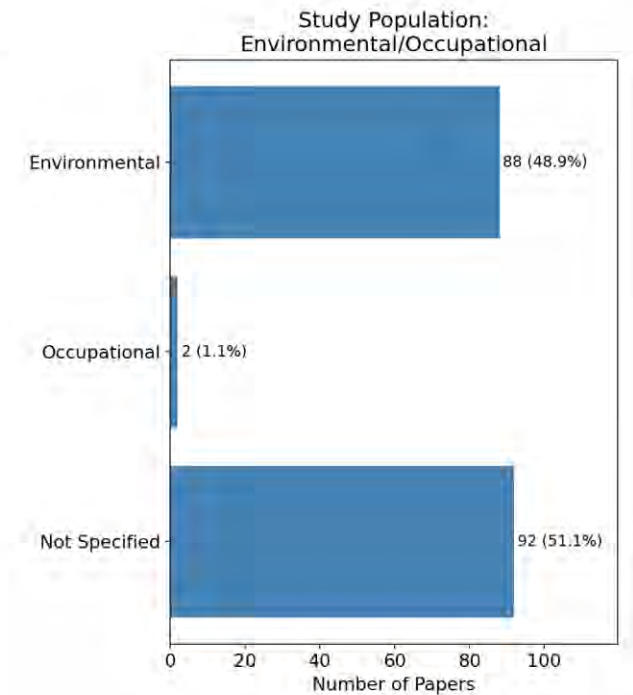
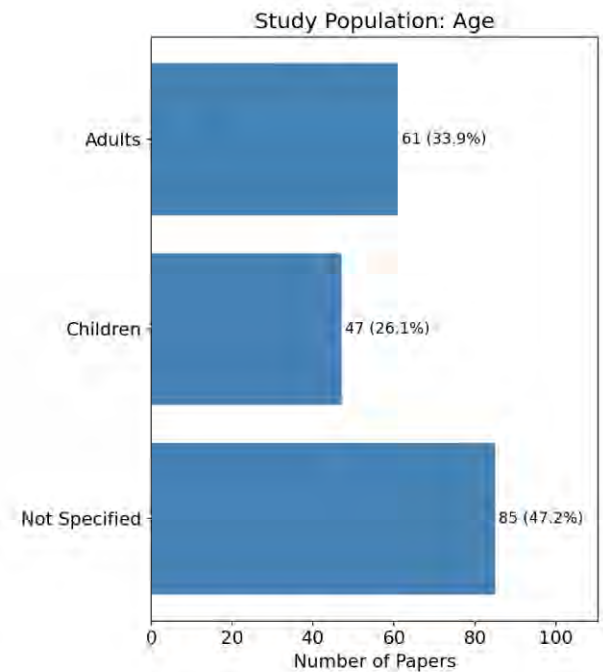
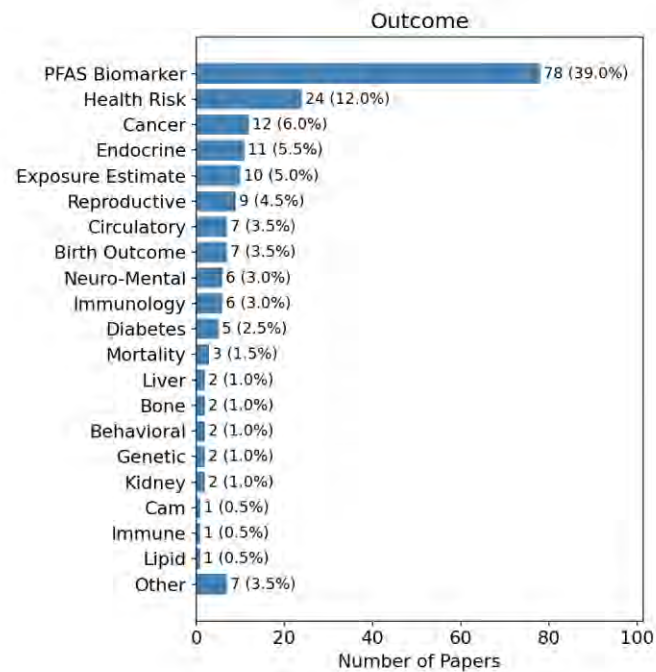
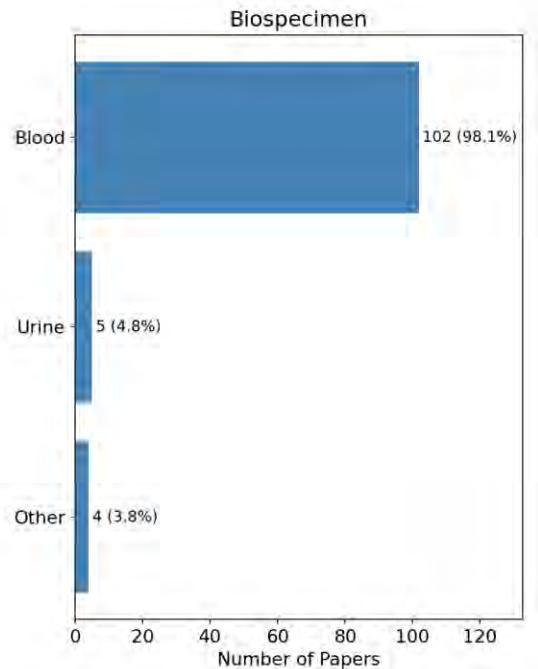
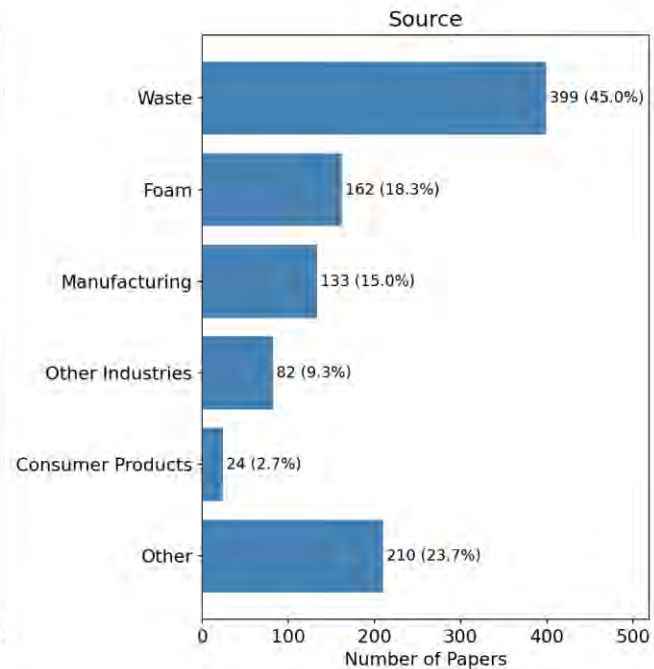
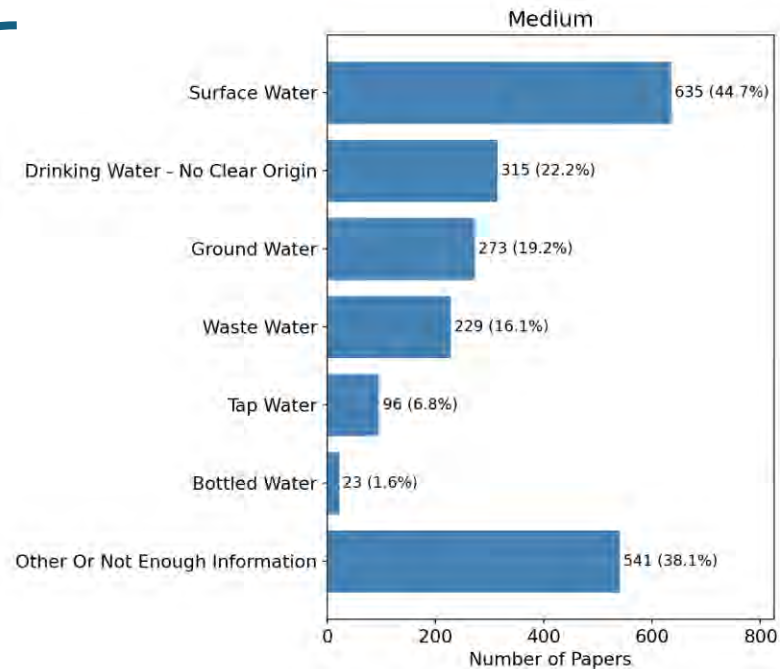
Entity Type	# Entities Evaluated	Match			# Incorrect Match	# No Match
		# Exact Match	# Partial Match	Total		
Environmental Medium	199	127 (63.8%)	32 (16.1%)	159 (79.9%)	35 (17.6%)	5 (2.5%)
Geographic Location	135	24 (17.8%)	78 (57.8%)	102 (75.6%)	30 (22.2%)	3 (2.2%)
Contamination Source	103	20 (19.4%)	60 (58.3%)	80 (77.7%)	11 (10.7%)	12 (11.7%)
Study Population	46	22 (47.8%)	11 (23.9%)	33 (71.7%)	8 (17.4%)	5 (10.9%)
Biospecimen	21	19 (90.5%)	0 (0.0%)	19 (90.5%)	1 (4.8%)	1 (4.8%)
Health/Environmental Outcome	46	21 (45.7%)	15 (32.6%)	36 (78.3%)	10 (21.7%)	0 (0.0%)
Overall	550	47.5%	31.5%	79%	15.7%	5.4%

Data analysis and visualization

Wu et al. (2022). Global performance and trends of research on per- and polyfluoroalkyl substances (PFASs) between 2001 and 2018 using bibliometric analysis. Chemosphere 295: 133853

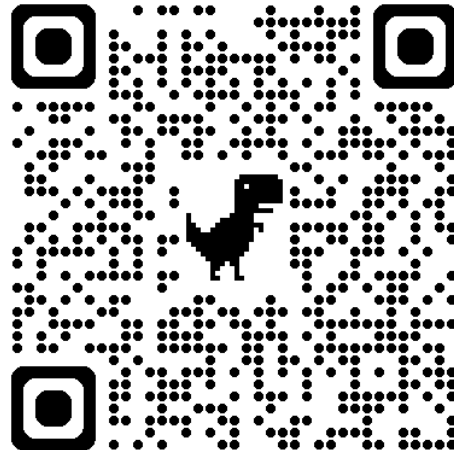


Distribution of entities



MedViz

<https://medviz.org/app/?dataset=medviz/pfas-2025-pfas-chemical>



MedViz: spaced-based exploration of biomedical literature



Embed articles into a high-dimensional semantic space, projects them to 2D, clusters them topically, and label them with LLMs



Visualize articles as an interactive semantic point cloud, where spatial proximity reflects semantic similarity



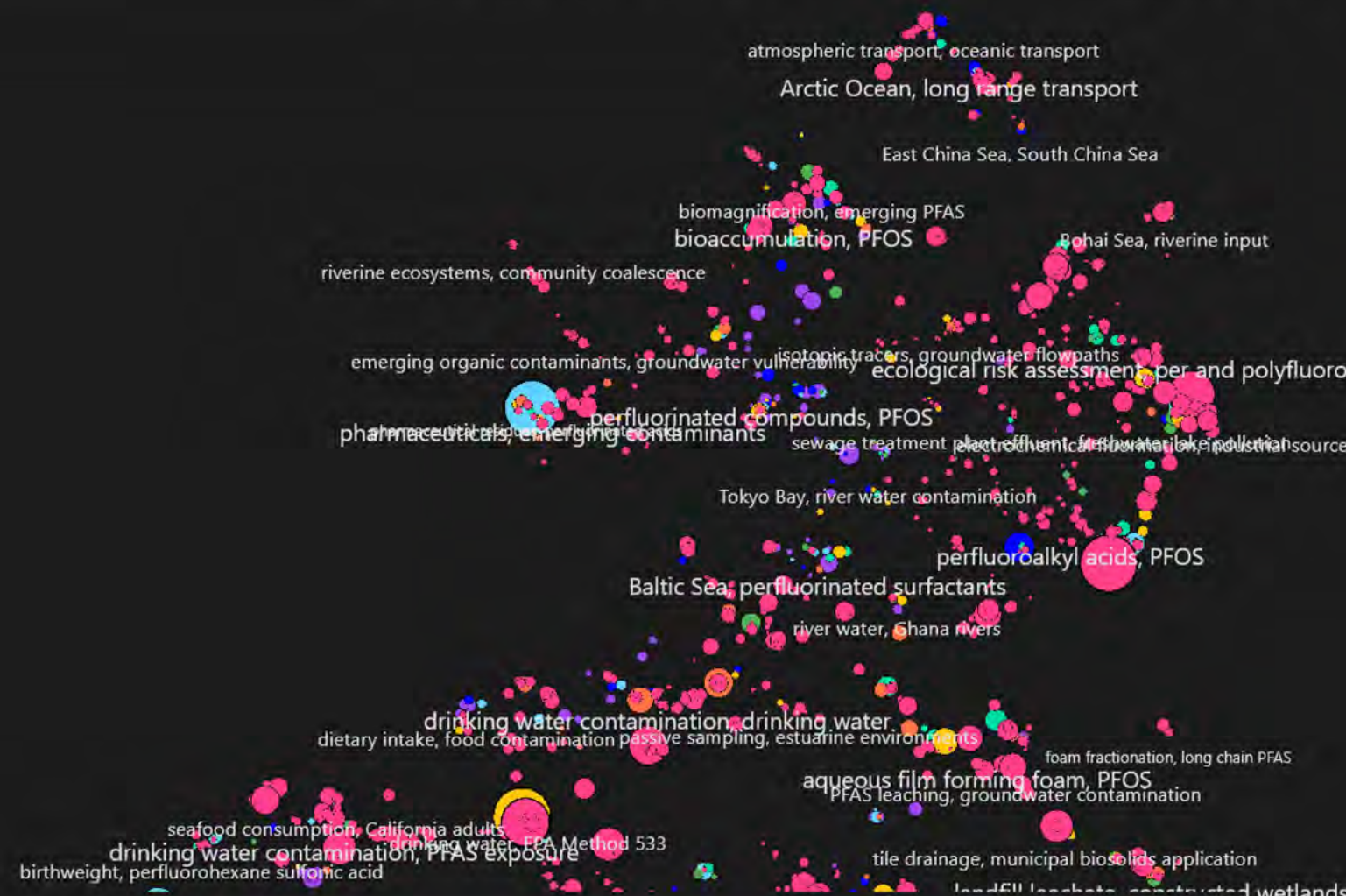
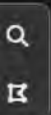
Explore temporal trends/patterns by filtering the same semantic space by publication year



Users can search, filter, and visually select articles to define explicit analytical context



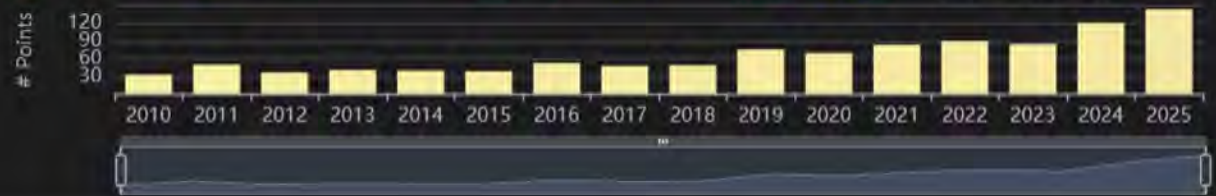
Visual selections provide context for agent-based question answering



Legend

- Before 2012
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018 - 2025

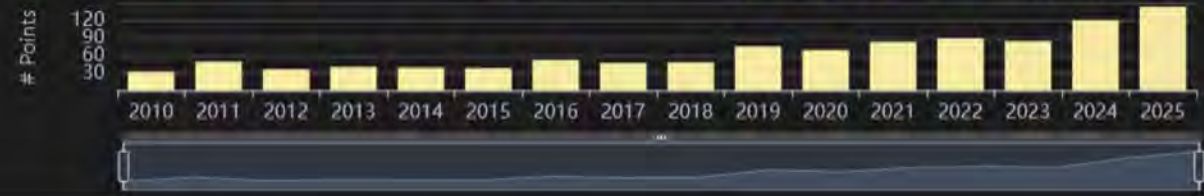
No Selection

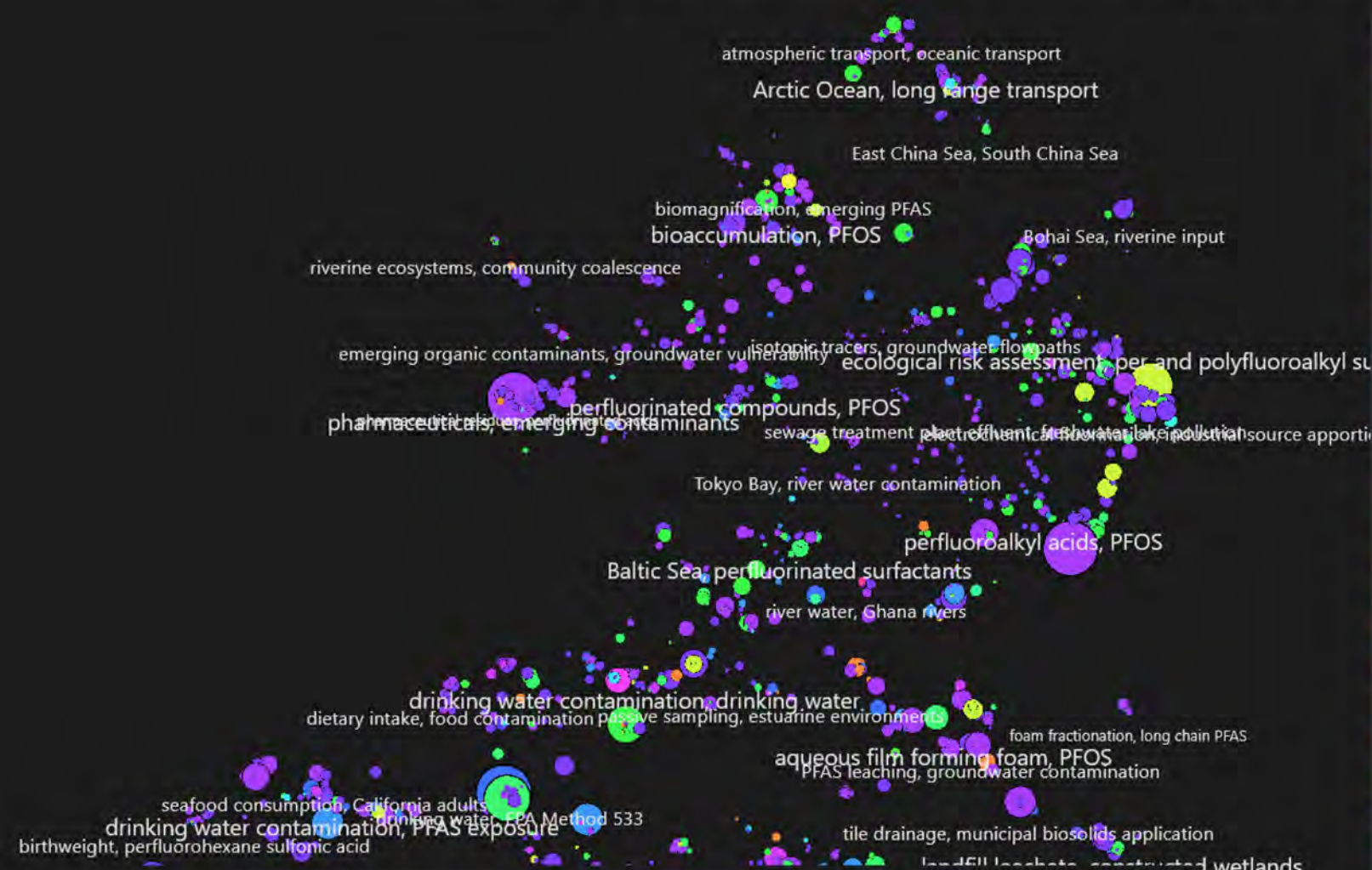


1,044 Papers 2010 - 2025



No Selection

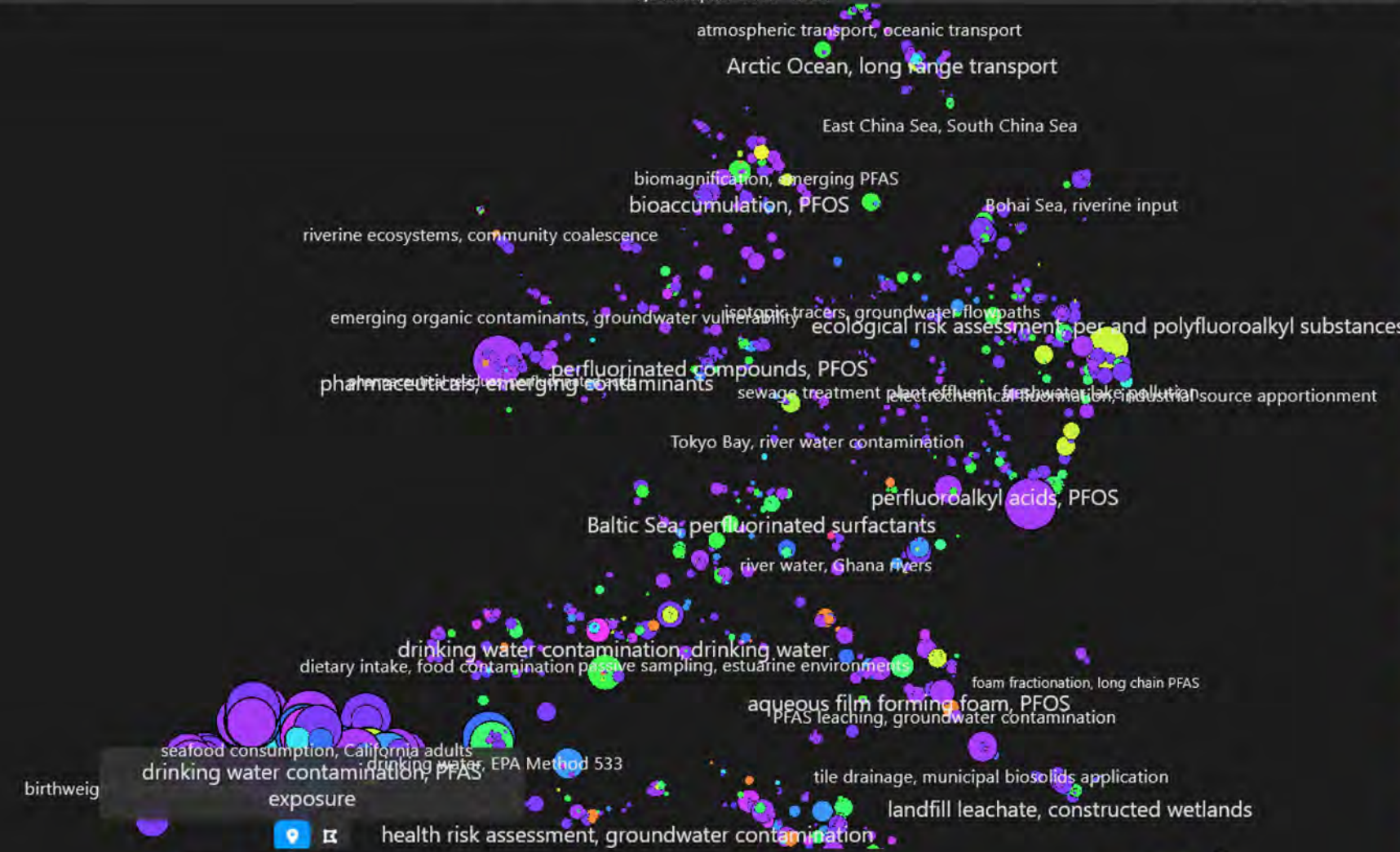




Legend

PFOA	376
PFOS	320
PFHxS	37
PFBS	62
PFBA	90
PFHxA	43
PFNA	12
PFHpA	15
PFPeA	29
6:2FTS	20
PFDA	3
HFPO-DA	37
PFUnA	1
PFDoA	1
NEtFOSAA	2
8:2FTS	1
PFPeS	1
NMeFOSAA	1
ADONA	1
PFHpS	1
PFTA	1
9CI-PF3ONS	1
PFTrDA	1
PFUnDA	1
4:2FTS	1
11CI-PF3OUdS	1
NFDHA	1
PFMPA	1





Legend

PF OA	376
PF OS	320
PF HxS	37
PF BS	62
PF BA	90
PF HxA	43
PF NA	12
PF HPA	15
PF PeA	29
6:2FTS	20
PF DA	3
HFPO-DA	17
PF UnA	1
PF DoA	
NE1FOSAA	2
8:2FTS	
PF PeS	1
NMeFOSAA	1
ADONA	2
PF HPS	
PFTA	
9CI-PF3ONS	1
PF TrDA	
PF UnDA	
4:2FTS	
11CI-PF3OUds	
NFDHA	
PF MPA	1

No Selection

Zoom in to this location on map







While PFOA and PFOS are well-known PFAS chemicals, "emerging" PFAS chemicals often refer to shorter-chain replacements like "GenX". Given this context, what are the emerging PFAS chemicals that are explicitly mentioned in this collection?

In the provided collection of research papers, several "emerging" PFAS chemicals are explicitly mentioned, particularly those that are shorter-chain alternatives to PFOA and PFOS. The identified emerging PFAS from the abstracts focus on:

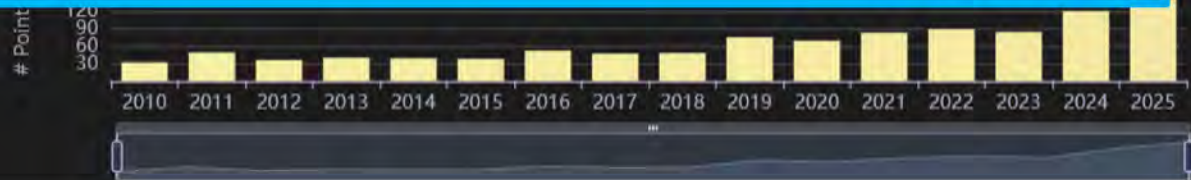
- Perfluorohexane sulfonate (PFHxS):** This chemical frequently appears across various studies, indicating its relevance in analyses related to exposure and health outcomes.
- Perfluorononanoic acid (PFNA):** Mentioned in the context of correlation with thyroid cancer incidence and other health impacts.
- Me-PFOSA:** Also referred to as a perfluoroalkyl substance analyzed in relation to eGFR (estimated Glomerular Filtration Rate) changes.

These substances, alongside PFOS and PFOA previously well-characterized, are increasingly important in discussions about PFAS impact on health and the environment. Specifically, they have been assessed for their effects on thyroid function, potential associations with diseases like cancer, and their prevalence in drinking water supplies.

While PFOA and PFOS are well-known PFAS chemicals, "emerging" PFAS chemicals often refer to shorter-chain replacements like "GenX". Given this context, what are the emerging PFAS chemicals that are explicitly mentioned in this collection?

Press **Enter** to send. **Shift + Enter** for a new line.

Record Send



1 Selection

Whole Space

30 | PFAS attributable c...

30 PFAS attributable cancer burden, Unreg...

Papers (30)

Prev 1 Next

Perfluorocarbon exposure, gender and thyroid function in the C8 Health Project. *The Journal of toxicological sciences* | 2011 | 21804304

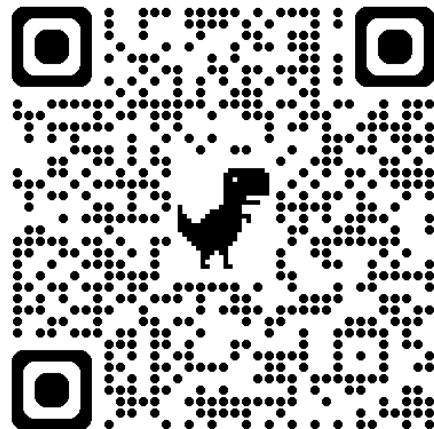
Perfluorocarbon exposure, gender and thyroid function in the C8 Health Project. *The Journal of toxicological sciences* | 2011 | 21804304

Associations between longitudinal serum perfluoroalkyl substance (PFAS) levels and measures of thyroid hormone, kidney function, and body mass index in the Fernald Community Cohort. *Environmental pollution (Barking, Essex : 1987)* | 2018 | 30373035

Associations between longitudinal serum perfluoroalkyl substance (PFAS) levels and measures of thyroid hormone, kidney function, and body mass

PFAS-Geomap

https://fair-ehs.med.yale.edu/pfas_map_slider_ns_leg_emerg.html



PFAS-Geomap: geographical exploration of PFAS exposure studies



Articles are visualized on an interactive world map



Temporal trends are explored by filtering the same space by publication year



Users can select an article (circle or square) to view its annotated abstract

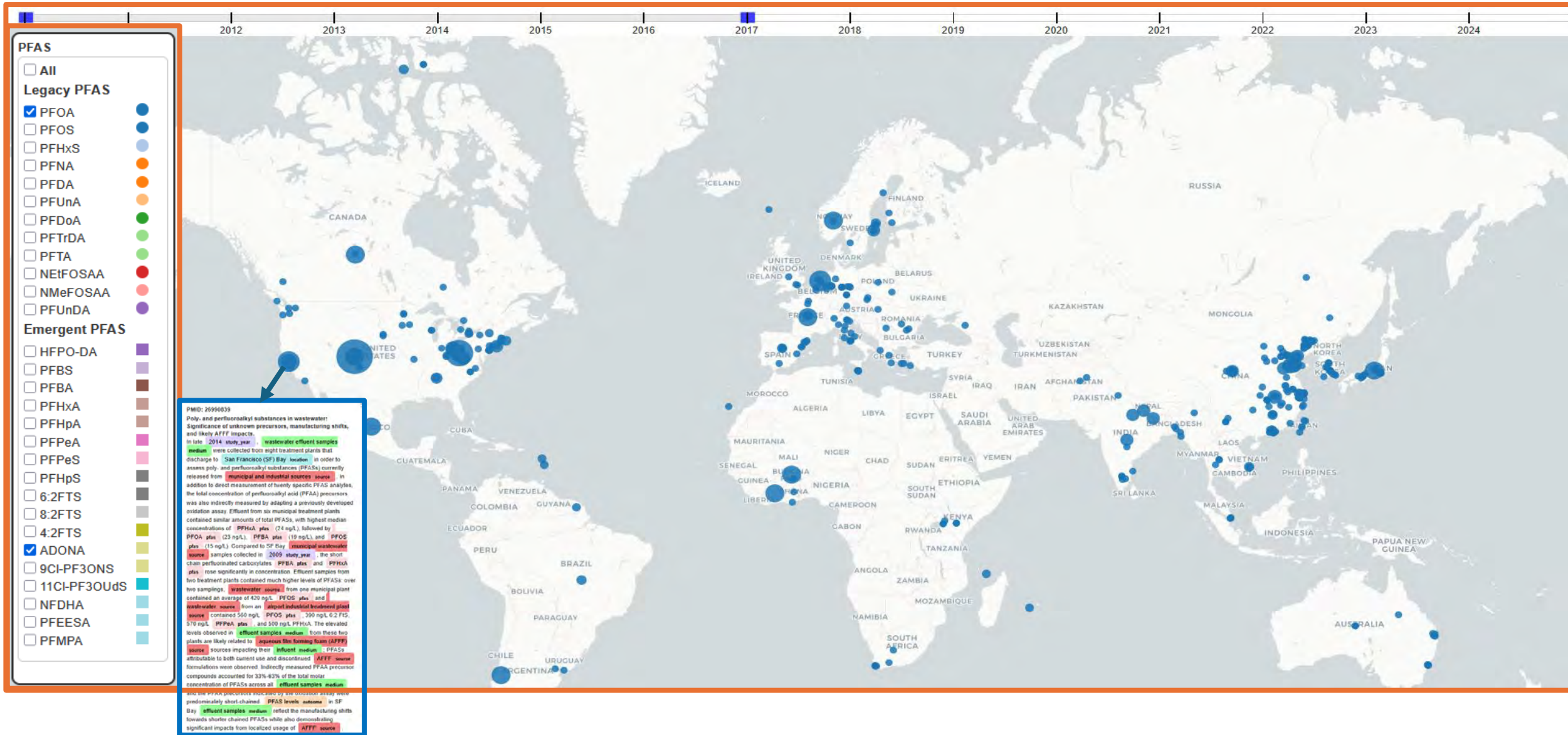


Google Map API is used to map geo-entities to geo-codes (latitude/longitude)

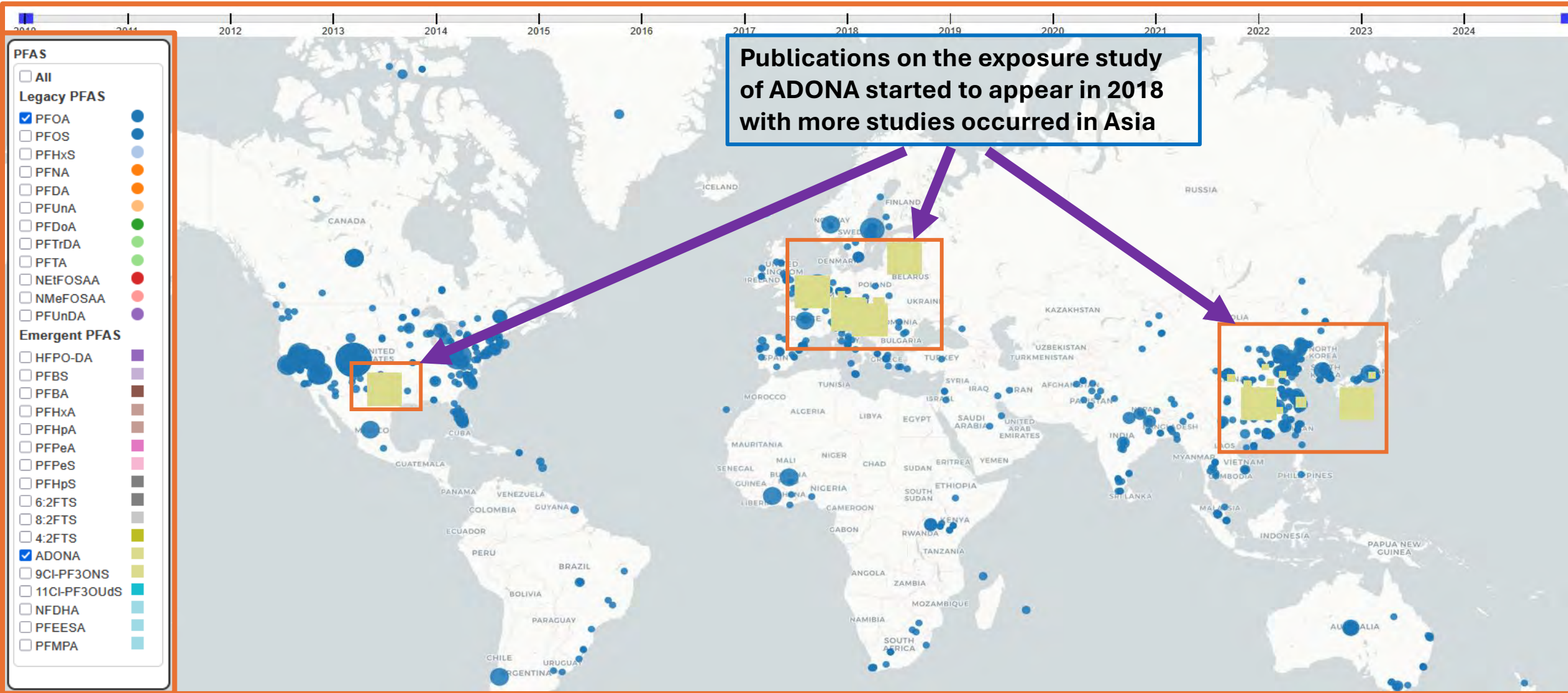


OpenStreetMap is used to implement the world map

PFOA and ADONA from 2010 to 2017



PFOA and ADONA from 2010 to 2025



Caveat



The visualization tools are primarily designed to facilitate exploration of extracted metadata



They help researchers identify patterns and trends that lead to the formulation of hypothesis-driven questions



They DO NOT act as a mechanism for drawing definitive research conclusions

Conclusions



Demonstrate the utility of an LLM-based approach for automated, scalable metadata extraction from environmental health literature



Broad applicability for accelerating evidence review and synthesis



Multidisciplinary collaboration is needed to ensure development of interpretable, robust, and human-centered AI systems



Our LLM approach can be applied to multiple environmental domains (e.g., air pollution, pesticide exposure)



Engage stakeholders across a wide spectrum of communities to broaden adoption of our LLM method

Future directions



Deeper metadata extraction from full-length papers



More detailed checklists including specialized templates



Use of LLM to aid in the development and refinement of templates



Enhance visualization



Integrate environmental data with EHR data



Expand user community engagement (e.g., All-Of-Us, exposomics)

Project Team

Biomedical Informatics & Data Science (Yale)

- Weiping Zhou
- Vipina Keloth
- Rixin Wang
- Haseena Rajeevan
- George Michel
- Casper Zhang
- Huan He
- Hua Xu

Superfund Research Program (Yale)

- Nicole Deziel
- Vasilis Vasiliou

CEDAR (Stanford)

- Mark Musen
- Martin O'Connor

EPA

- Elaine Hubal

Acknowledgments

NIEHS grant R24ES036135



Thanks!