

# Extraction of chemical structures from literature and patent documents using open access chemistry toolkits: a case study with PFAS

**Assoc. Prof. Emma L. Schymanski**

Environmental Cheminformatics, Luxembourg Centre for  
Systems Biomedicine, University of Luxembourg

Contact: [emma.schymanski@uni.lu](mailto:emma.schymanski@uni.lu) Twitter: [@ESchymanski](https://twitter.com/ESchymanski)

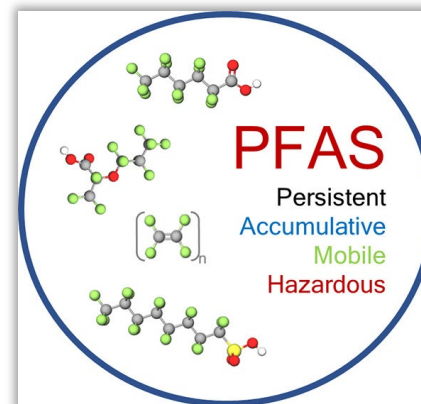
Shadrack J Barnabas<sup>1</sup>, Timo Böhme<sup>1</sup>, Stephen K Boyer<sup>2</sup>,  
Matthias Irmer<sup>1</sup>, Christoph Ruttkies<sup>1</sup>, Ian Wetherbee<sup>3</sup>, Todor Kondic<sup>4</sup>, Lutz Weber<sup>1</sup>  
<sup>1</sup>OntoChem GmbH, <sup>2</sup>Collabra Inc., <sup>3</sup>Google LLC, <sup>4</sup>LCSB, University of Luxembourg

	Definition A	Definition B	Definition C
# PFAS in:			
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197



# Outline for today

- Motivation
  - What are PFAS – and why?
- Extracting Chemical Data
- Different Toolkits and Definitions
- Helping People Find PFAS
- Questions / Discussion



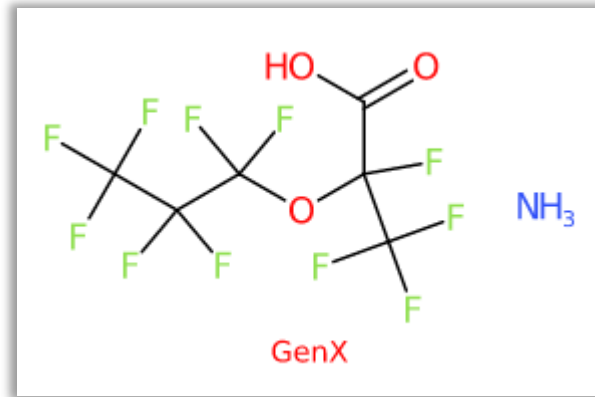
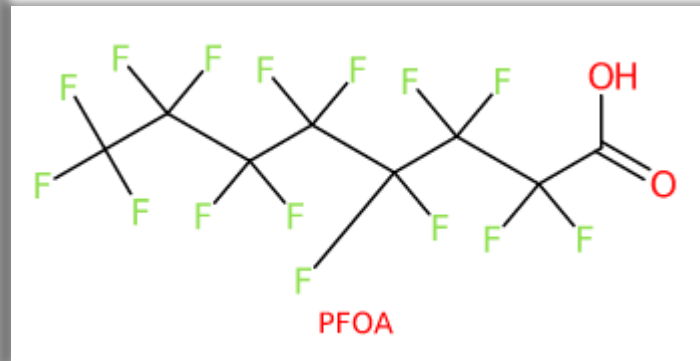
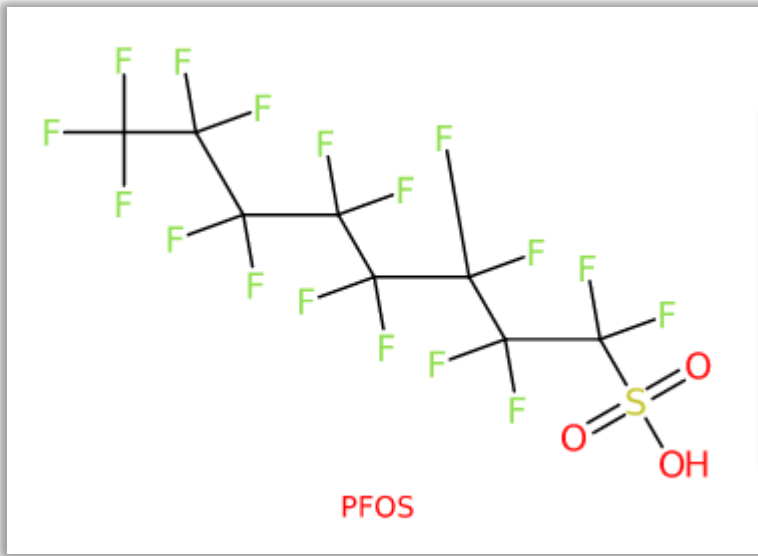
	Definition A	Definition B	Definition C
# PFAS in:			
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197

The table is accompanied by three chemical structures representing different PFAS definitions. Definition A shows a central carbon atom bonded to two R groups (R<sup>1</sup>, R<sup>2</sup>) and two fluorine atoms (F). Definition B shows a central carbon atom bonded to two AH groups and two fluorine atoms (F). Definition C shows a central carbon atom bonded to two R groups (R<sup>1</sup>, R<sup>2</sup>) and two fluorine atoms (F).

OntoChem PFAS lists	Count
OntoChem PFAS from CORE - Definition A	26,805
OntoChem PFAS from CORE - Definition B	4,114
OntoChem PFAS from CORE - Definition C	3,432
OntoChem PFAS from Google Patents - Definition A	1,762,971
OntoChem PFAS from Google Patents - Definition B	73,749
OntoChem PFAS from Google Patents - Definition C	33,649

# What are PFAS?

PFAS = Per- and polyfluoroalkyl substances



<https://defence.gov.au/environment/pfas/PFAS.asp>

<https://www.defence.gov.au/copyright>

<https://chemtrust.org/dark-waters-film-pfas/>

Google

ontochem

uni.lu

LCSB

# Why PFAS?



## Scientific Basis for Managing PFAS as a Chemical Class

Carol F. Kwiatkowski\*, David Q. Andrews, Linda S. Birnbaum, Thomas A. Bruton, Jamie C. DeWitt, Detlef R. U. Knappe, Maricel V. Maffini, Mark F. Miller, Katherine E. Pelch, Anna Reade, Anna Soehl, Xenia Trier, Marta Venier, Charlotte C. Wagner, Zhanyun Wang, and Arlene Blum

✓ Cite this: *Environ. Sci. Technol. Lett.* 2020, 7, 8, 532–543

Publication Date: June 30, 2020

<https://doi.org/10.1021/acs.estlett.0c00255>

Copyright © 2020 American Chemical Society

[RIGHTS & PERMISSIONS](#)

Article Views

50437

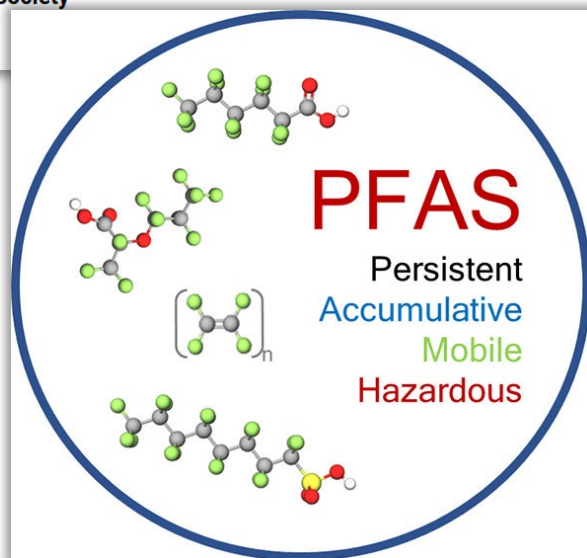
Altmetric

417

Citations

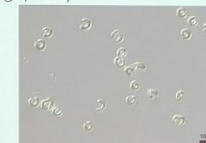
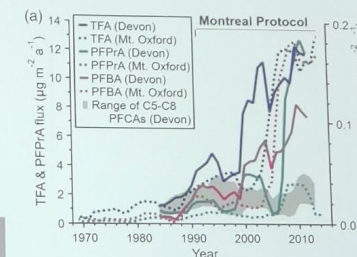
132

[LEARN ABOUT THESE METRICS](#)



## There is no effective dilution to persistent global pollution

- Ice core records show accumulation of TFA and other short-chain PFAS; all evidence points to anthropogenic origin
- Lowest no-observable effect concentration so far: *Raphidocelis subcapitata* (120 µg/L)
- If remote levels reach threshold concentration at remote regions, there is no way of reversing quickly



<https://alchetron.com/Raphidocelis-subcapitata>

Pickard et al., *Geophysical Research Letters* (2020), 47, e2020GL087535  
Jouanda (2021), *ESPI* 23(11), 1641-1649.  
Bouttonnet et al., *Hum Ecol Risk Assess.* 1999;5:59-124.

SETAC Europe 32<sup>nd</sup> Annual Meeting

#SETACCopenhagen

# Zero PM

Prevent

Prioritize

Remove

@annaruthski found high levels of Nafion BP2 (a Chemours PFAS) in the livers of dead seabirds along the Cape Fear River estuaries—which is over 80 miles from the Fayetteville facility. 2/



[pulse.ncpolicywatch.org](https://pulse.ncpolicywatch.org)

Toxic PFAS found in livers of Atlantic seabirds, including those in NC | The Pulse Per- and polyfluoroalkyl substances, or PFAS, are accumulating in dead juvenile seabirds in three locations off the Atlantic Coast, including North Carolina, ...

<https://twitter.com/CleanCapeFear/status/1519659268571602944>

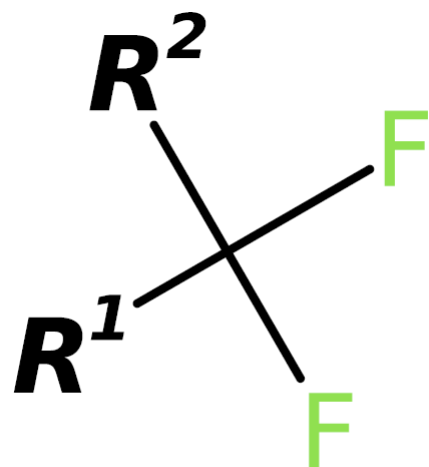
<https://twitter.com/ESchymanski/status/1530958689720467457/>



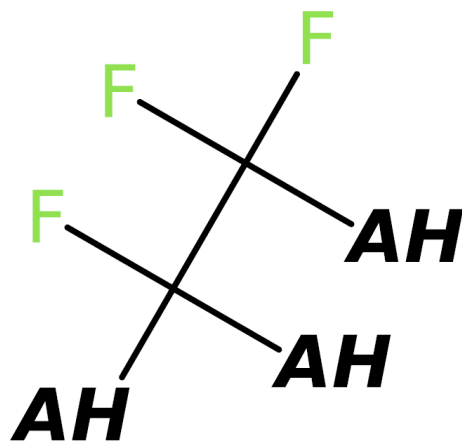
Kwiatkowski et al. (2020) ES&TL DOI: [10.1021/acs.estlett.0c00255](https://doi.org/10.1021/acs.estlett.0c00255)  
Schymanski (2022) GRC Env.Sci.Wat. DOI: [10.5281/zenodo.6655802](https://doi.org/10.5281/zenodo.6655802)

# Definitions of PFAS

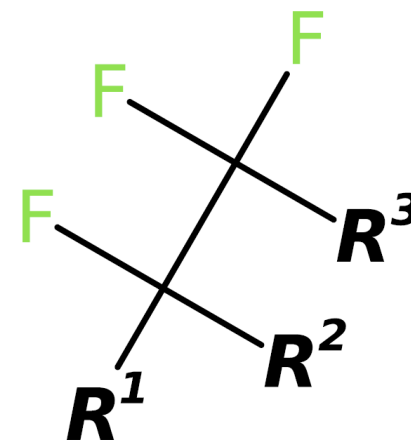
There are many definitions out there ...



Definition A  
"OECD PFAS"



Definition B



Definition C

AH = any atom, including H  
 $R^n$  = any atom, except H

...

Google

ontochem

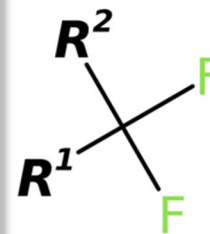


# The Scale of the PFAS Problem...

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120>



- ▼ PFAS and Fluorinated Compounds in PubChem ? ↗ 19,692,732
  - ▼ OECD PFAS definition ? ↗ 6,203,598
    - ▶ Molecule contains isolated CF2 ? 598,959
    - ▶ Molecule contains isolated CF3 ? 5,485,764
    - ▶ Molecule contains PFAS parts larger than CF2/CF3 ? 222,650
  - ▼ Organofluorine compounds ? ↗ 19,543,301
    - ▶ Fluorinated aliphatic substances ? 850,949
    - ▶ Fluorinated aromatic substances ? 18,630,145
    - ▶ Other fluorinated substances ? 84,042
  - ▼ Other diverse fluorinated compounds ? 114,526
    - ▶ Contains fluorine bond to non-carbon element ? 23,457
    - ▶ Contains non-organic element ? 113,455
  - ▶ PFAS and fluorinated compound collections ? ↗ 1,785,540



>6 million  
OECD PFAS  
in PubChem

# The Scale of the PFAS Problem...

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120>



PFAS and Fluorinated Compounds in PubChem **19,692,732**

OECD PFAS definition **6,203,598**

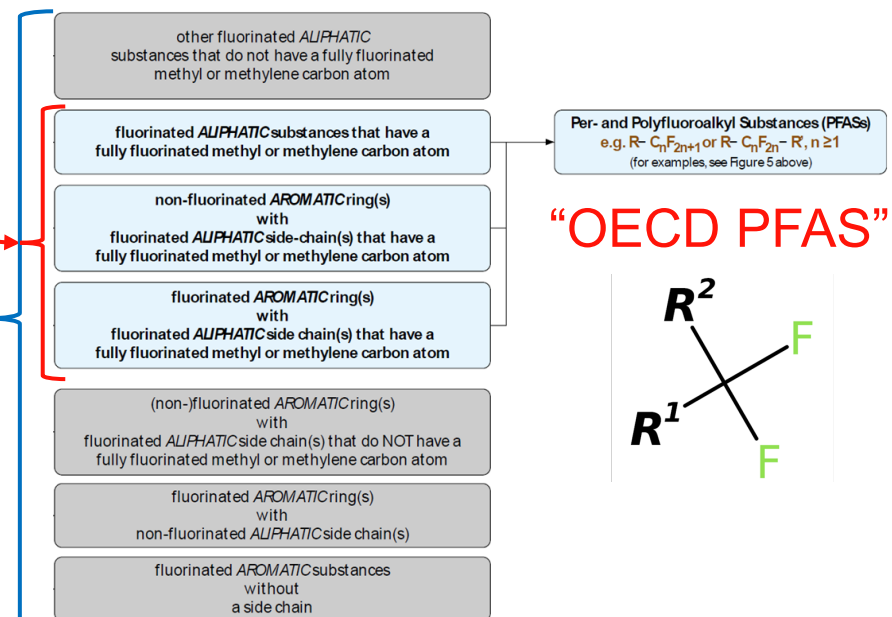
- ▶ Molecule contains isolated CF<sub>2</sub> **598,959**
- ▶ Molecule contains isolated CF<sub>3</sub> **5,485,764**
- ▶ Molecule contains PFAS parts larger than CF<sub>2</sub>/CF<sub>3</sub> **222,650**

Organofluorine compounds **19,543,301**

- ▶ Fluorinated aliphatic substances **850,949**
- ▶ Fluorinated aromatic substances **18,630,145**
- ▶ Other fluorinated substances **84,042**

Other diverse fluorinated compounds **114,526**

- ▶ Contains fluorine bond to non-carbon element **23,457**
- ▶ Contains non-organic element **113,455**
- ▶ PFAS and fluorinated compound collections **1,785,540**



**>6 million  
OECD PFAS  
in PubChem**

# The Scale of the PFAS Problem vs Existing PFAS Lists



▼ PFAS and Fluorinated Compounds in PubChem ? ↗ 19,692,732

▼ OECD PFAS definition ? ↗ 6,203,598

▶ Molecule contains isolated CF2 ? 598,959

▶ Molecule contains isolated CF3 ? 5,485,764

▶ Molecule contains PFAS parts larger than CF2/CF3 ?

▼ Organofluorine compounds ? ↗ 19,543,301

▶ Fluorinated aliphatic substances ? 850,949

▶ Fluorinated aromatic substances ? 18,630,145

▶ Other fluorinated substances ? 84,042

▼ Other diverse fluorinated compounds ? 114,526

▶ Contains fluorine bond to non-carbon element ? 23,457

▶ Contains non-organic element ? 113,455

▶ PFAS and fluorinated compound collections ? ↗ 1,785,540

▼ PFAS and fluorinated compound collections ? ↗ 1,785,540

▶ CompTox Chemicals Dashboard PFAS suspect lists ? ↗ 8,502

▶ NORMAN-SLE PFAS suspect lists ? ↗ 6,040

▶ OntoChem PFAS lists ? ↗ 1,777,052

▶ Other fluorinated chemical content in PubChem ? ↗ 1,674

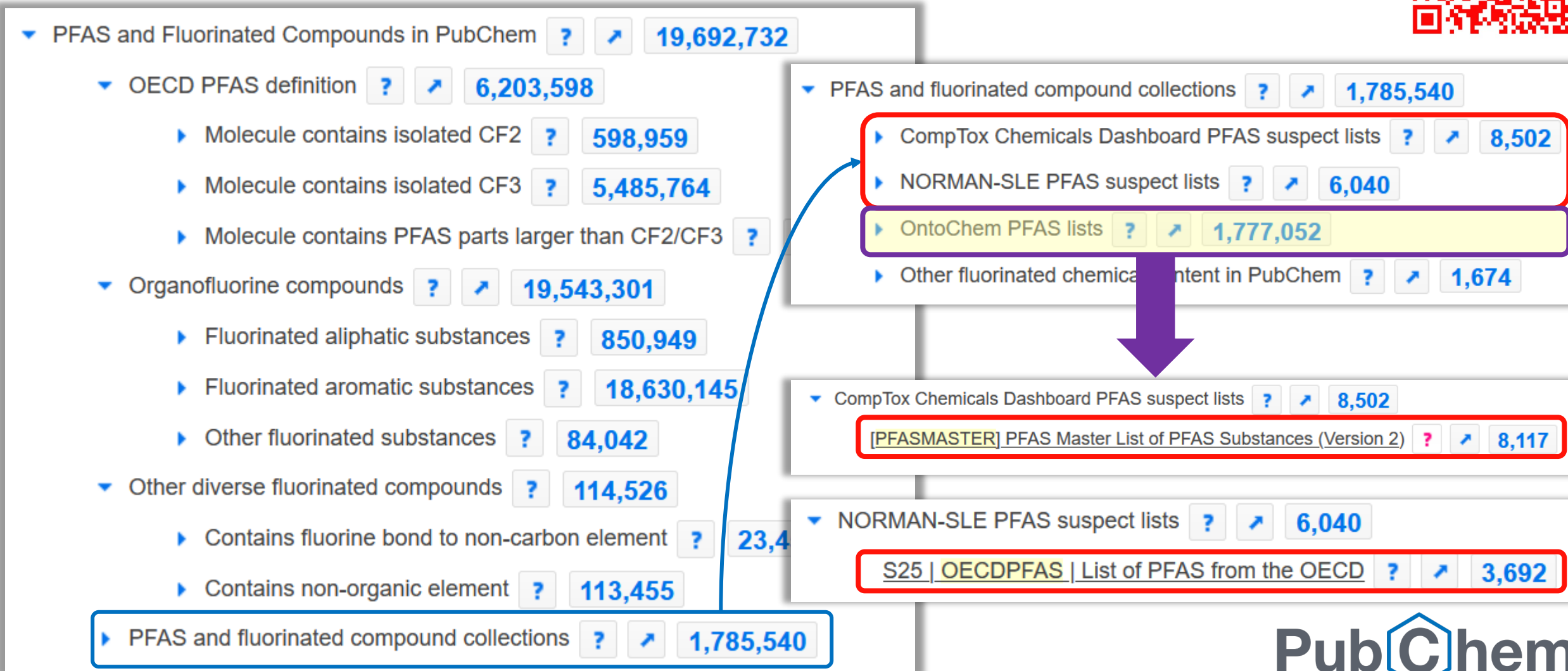
PubChem

NIH U.S. National Library of Medicine  
National Center for Biotechnology Information





# The Scale of the PFAS Problem vs Existing PFAS Lists



Schymanski & Bolton (2022) POPRC18 Event. DOI: [10.5281/zenodo.7118552](https://doi.org/10.5281/zenodo.7118552)

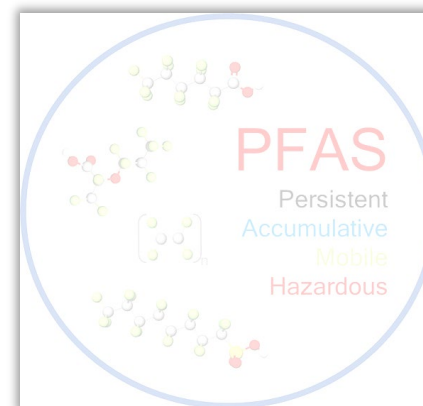
<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120>

Barnabas *et al.* (2022) *Digital Discovery*. DOI: [10.1039/D2DD00019A](https://doi.org/10.1039/D2DD00019A)



# Outline for today

- Motivation
  - What are PFAS – and why?
- Extracting Chemical Data
- Different Toolkits and Definitions
- Helping People Find PFAS
- Questions / Discussion



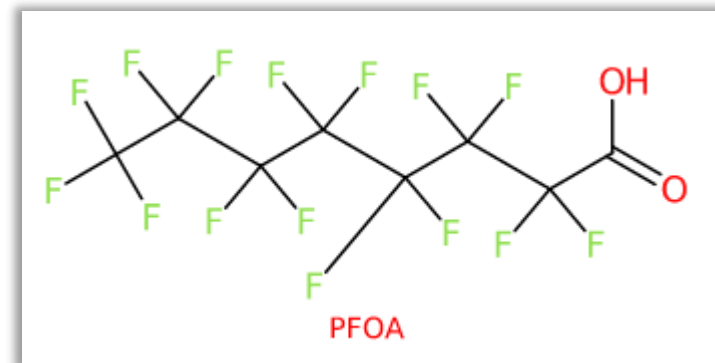
# PFAS in:	Definition A	Definition B	Definition C
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197

OntoChem PFAS lists	1,777,052
OntoChem PFAS from CORE - Definition A	26,805
OntoChem PFAS from CORE - Definition B	4,114
OntoChem PFAS from CORE - Definition C	3,432
OntoChem PFAS from Google Patents - Definition A	1,762,971
OntoChem PFAS from Google Patents - Definition B	73,749
OntoChem PFAS from Google Patents - Definition C	33,649

# Challenges Extracting Chemical Data - Names

PFOA

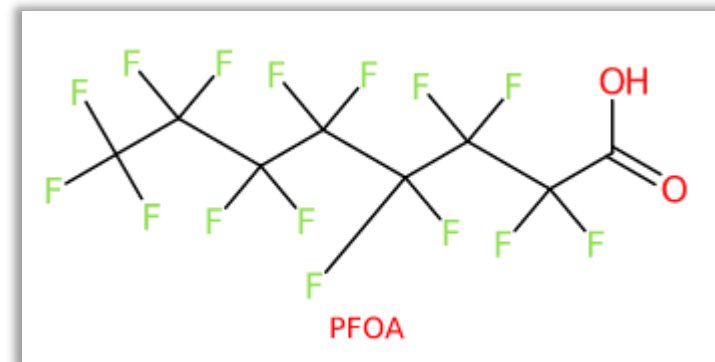
Perfluorooctanoic acid



# Challenges Extracting Chemical Data - Names

## PFOA

## Perfluorooctanoic acid



**PubChem** Perfluorooctanoic acid (Compound)

### 2.4.2 Depositor-Supplied Synonyms



PERFLUOROOCTANOIC ACID	Pentadecafluoro-1-octanoic acid	Hexanoyl fluoride,
Pentadecafluorooctanoic acid	Pentadecafluoro-n-octanoic acid	947VD76D3L
335-67-1	Perfluoro-n-octanoic acid	MFCD00004174
PFOA	IPC-PFFA-8 HG	Octanoic acid, 2,2,
Perfluorocaprylic acid	NSC 95114	CCRIS 4386
2,2,3,3,4,4,5,5,6,6,7,7,8,8-Pentadecafluorooctanoic acid	UNII-947VD76D3L	HSDB 7137
Octanoic acid, pentadecafluoro-	n-perfluorooctanoic acid	EINECS 206-397-9
Perfluorooctanoic acid	DTXSID8031865	IPC-PFFA-8
Perfluoroheptanecarboxylic acid	CHEBI:35549	BRN 1809678

79 synonyms  
in **PubChem**!

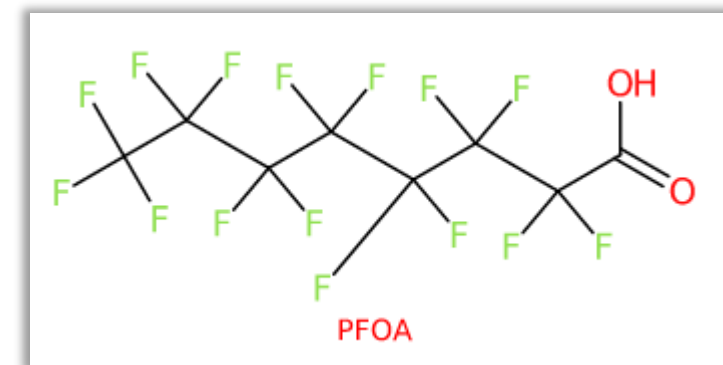
<https://pubchem.ncbi.nlm.nih.gov/compound/9554#section=Depositor-Supplied-Synonyms>

**ontochem**

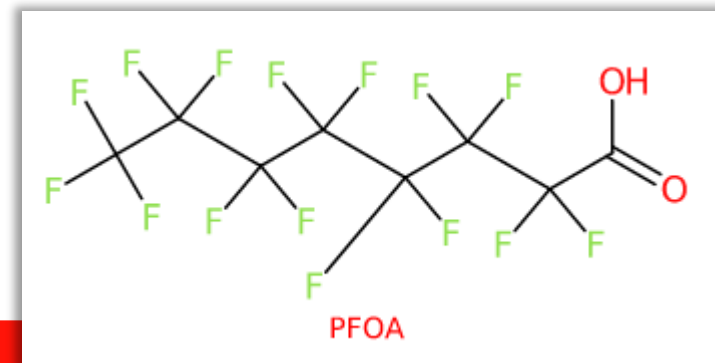


# Challenges Extracting Chemical Data - Identifiers

Database	Identifier
CAS Registry	335-67-1
EC Number	206-397-9
OntoChem	OCID190000019728
PubChem	9554
CompTox	DTXSID8031865
ChemSpider	9180

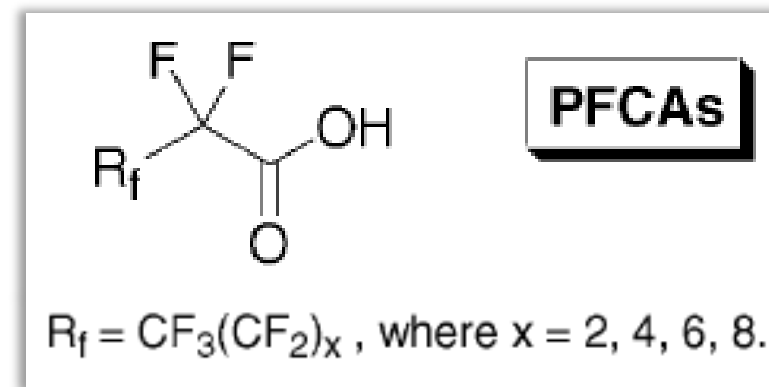
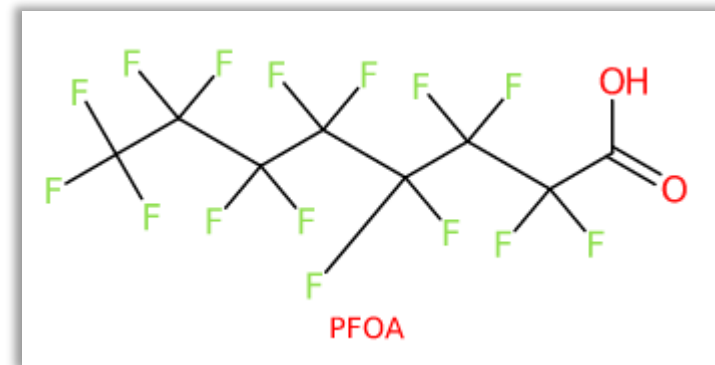
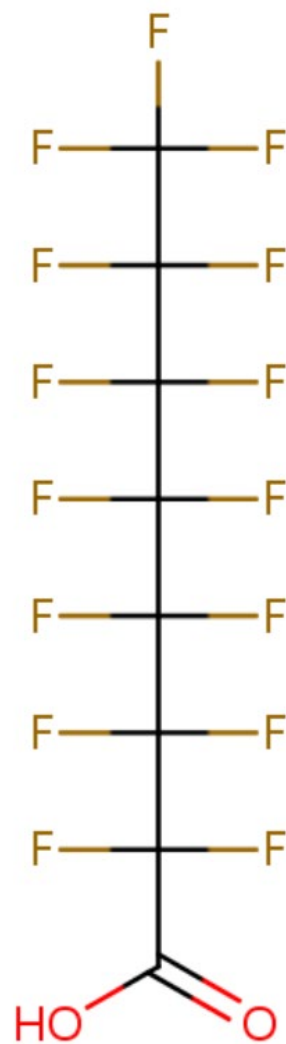
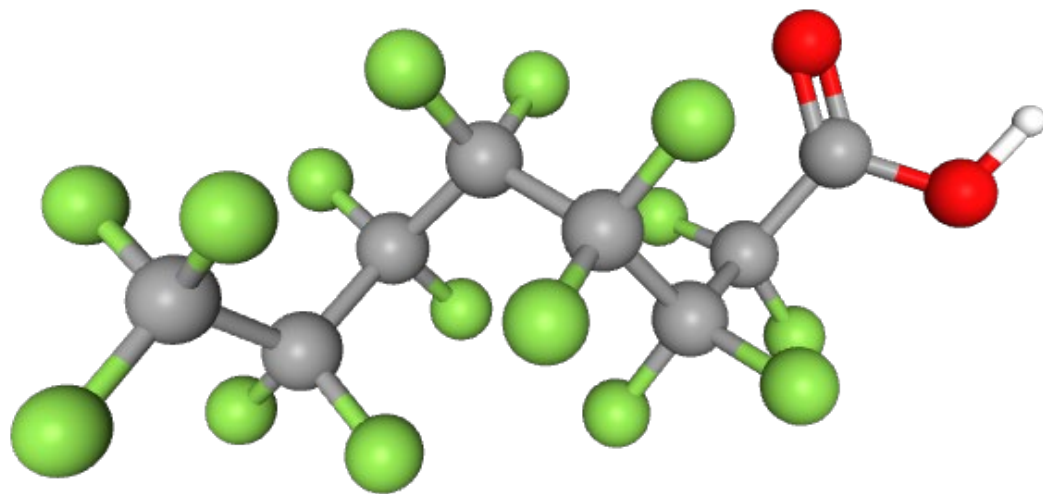
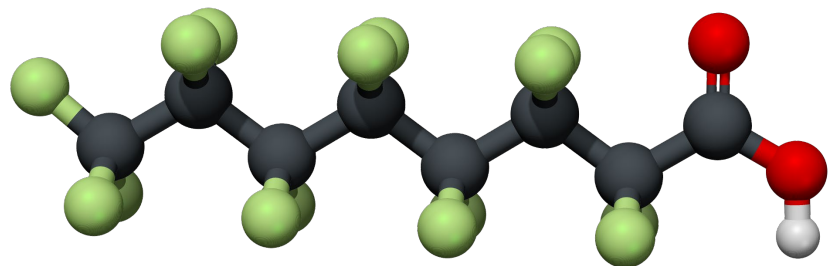


# Challenges Extracting Chemical Data - Cheminf. Formats

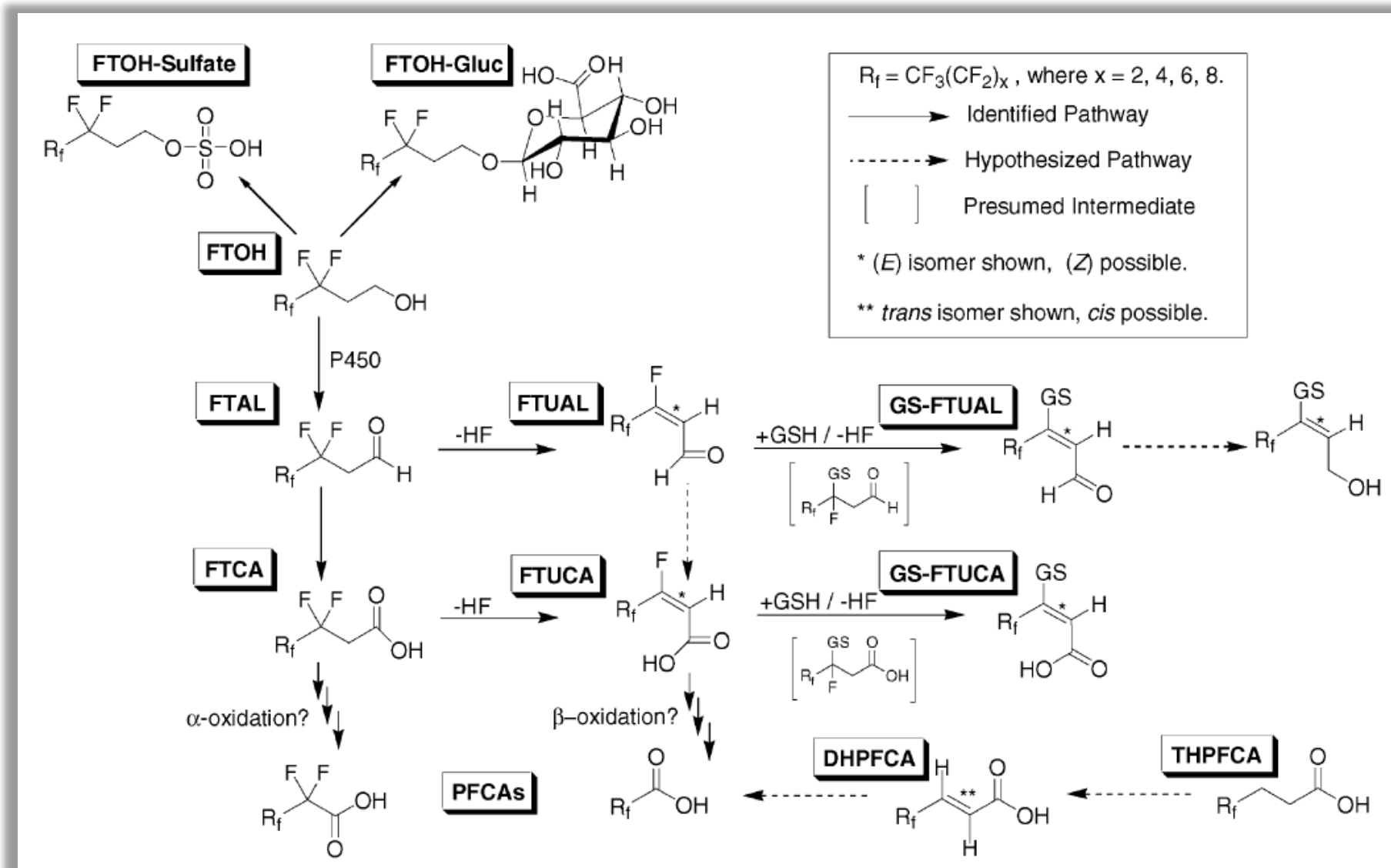


Format	Description	Example
SMILES	Simplified Molecular Input Line Entry System	<chem>C(=O)(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O</chem>
InChI	International Chemical Identifier	InChI=1S/ <chem>C8HF15O2</chem> /c9-2(10,1(24)25)3(11,12)4(13,14)5(15,16)6(17,18)7(19,20)8(21,22)23/h(H,24,25)
InChIKey	Hashed InChI	SNGREZUHAYWORS- <chem>UHFFFAOYSA</chem> -N

# Challenges Extracting Chemical Data - Images



# Challenges Extracting Chemical Data - Images





Datasets Used: CORE Repository: <https://core.ac.uk/>

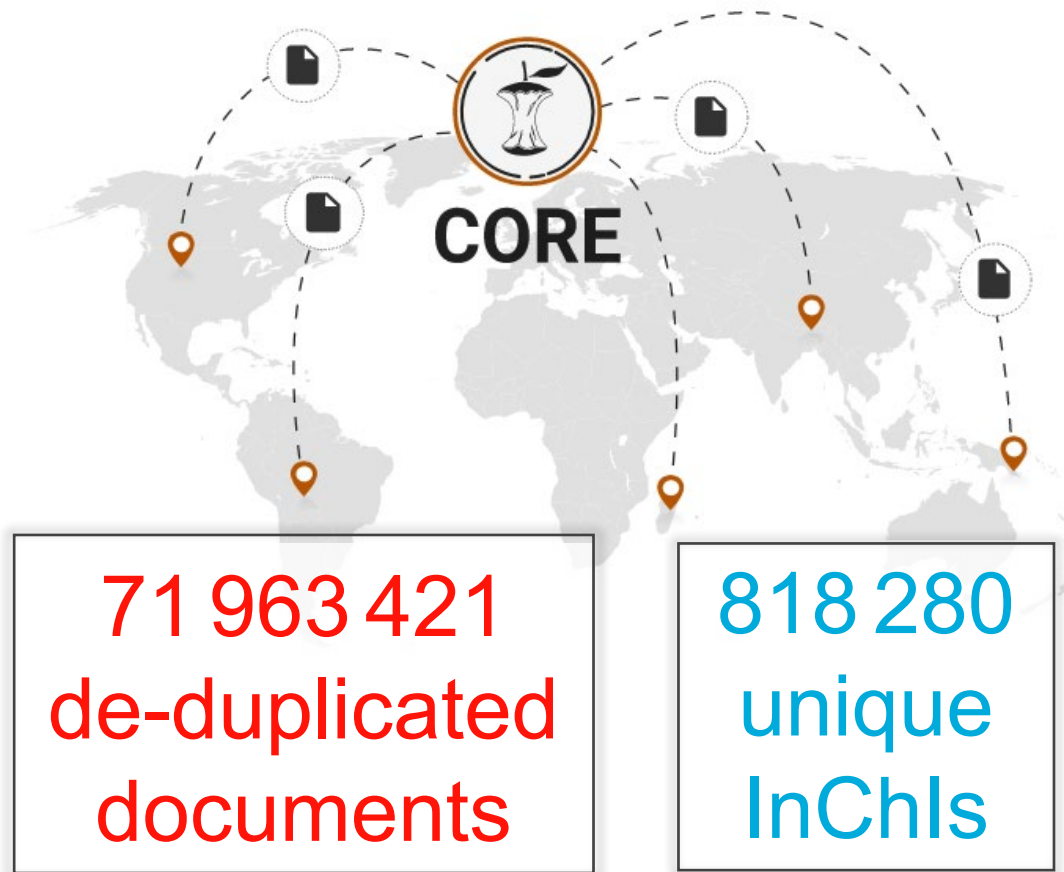


Services ▾ Communities ▾ About ▾

# The world's largest collection of open access research papers

Search 232M from papers around the world

SEARCH



# Datasets Used: Google Patent Corpus

The image shows a composite screenshot of two web interfaces. On the left is the Google Patents search page, and on the right is the Google Cloud Explorer interface for a BigQuery table.

**Google Patents Interface:**

- Search bar: |
- SEARCH TERMS: + Synonym
- SEARCH FIELDS: Date · Priority (YYYY-MM-DD - YYYY-), + Inventor, + Assignee, Patent Office, Language, Status, Type, Litigation.

**Google Cloud Explorer Interface:**

- Project: Select a project
- Search: Products, resources, docs (/)
- Explorer: Viewing all resources. [Show starred resources only.](#)
- Resources: sciwalker-open-data, patents-public-data
- Table: annotations\_202101
- Info: This is a partitioned table. [Learn more](#)
- Schema Table:

Filter	Field name	Type	Mode	Collation	Default Value	Policy Tags
<input type="checkbox"/>	<a href="#">publication_number</a>	STRING	NULLABLE			
<input type="checkbox"/>	<a href="#">ocid</a>	INTEGER	NULLABLE			
<input type="checkbox"/>	<a href="#">preferred_name</a>	STRING	NULLABLE			

# Datasets Used: Google Patent Corpus

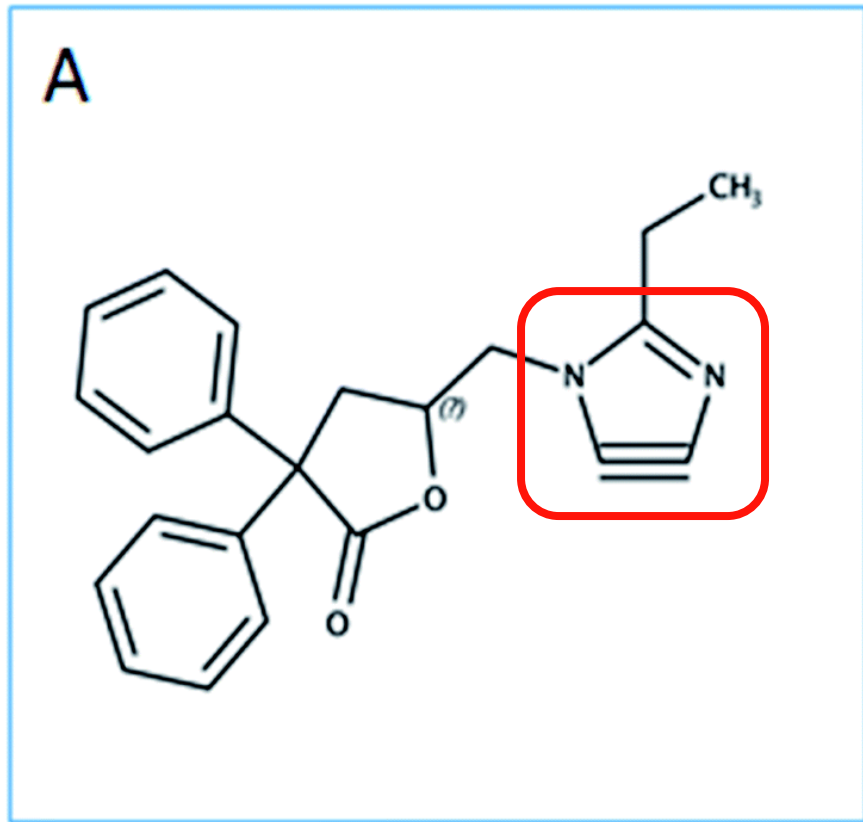
The image shows a screenshot of the Google Patents website with several red callout boxes highlighting key statistics. In the background, a Google Cloud console window is open, showing a table schema for a dataset. The statistics are as follows:

- 51 928 230 588 annotations** (Total annotations)
- 18 032 261 unique InChIs** (Unique InChI strings)
- 4 533 988 229 compound annotations** (Total compound annotations)
- 4 182 712 F-containing SMILES** (F-containing SMILES)

The background Google Cloud console shows a table with the following schema:

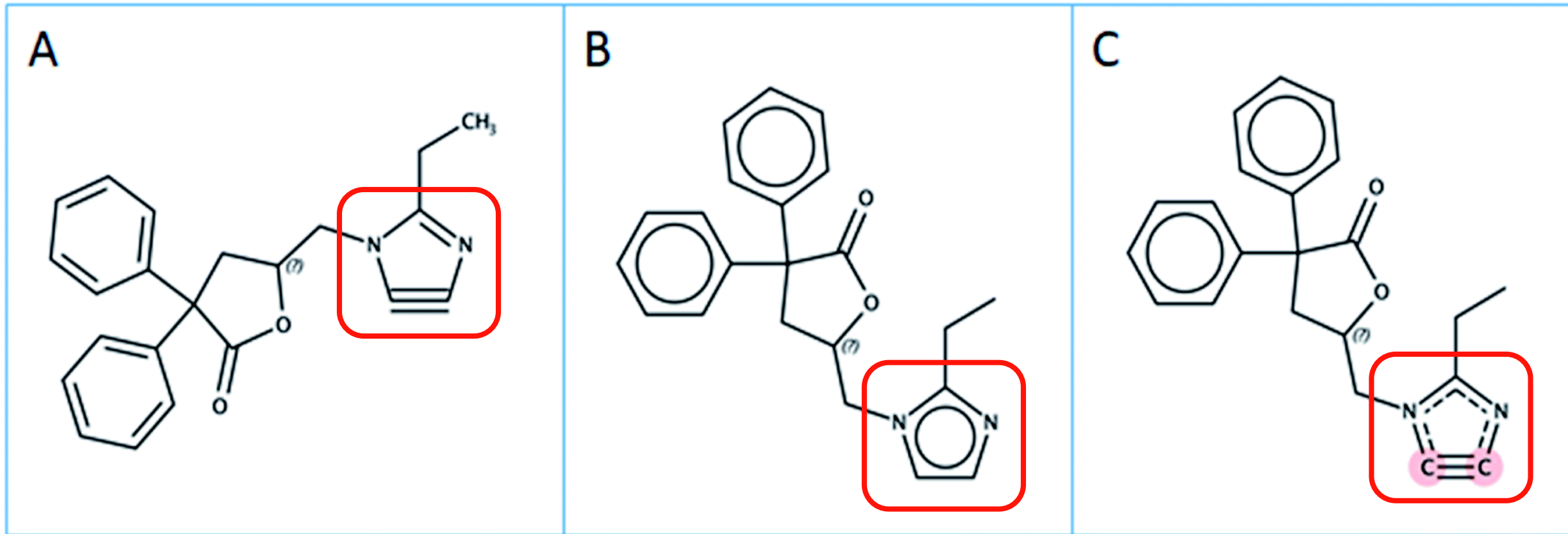
Field name	Type	Mode
<a href="#">publication_number</a>	STRING	NULLABLE
<a href="#">ocid</a>	INTEGER	NULLABLE
<a href="#">preferred_name</a>	STRING	NULLABLE

# Challenges Extracting Chemical Data – Toolkit Differences



OSRA Input SMILES

# Challenges Extracting Chemical Data – Toolkit Differences



ORSA Input SMILES

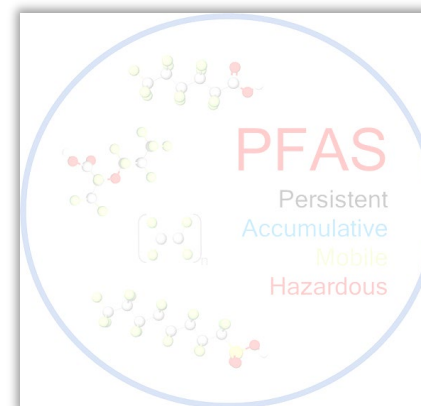
ChemAxon & OCL  
interpretation

RDKit interpretation

Rejected by CDK

# Outline for today

- Motivation
  - What are PFAS – and why?
- Extracting Chemical Data
- Different Toolkits and Definitions
- Helping People Find PFAS
- Questions / Discussion



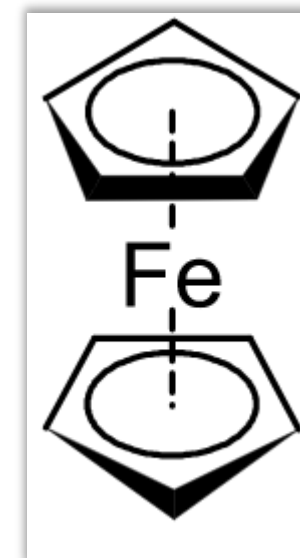
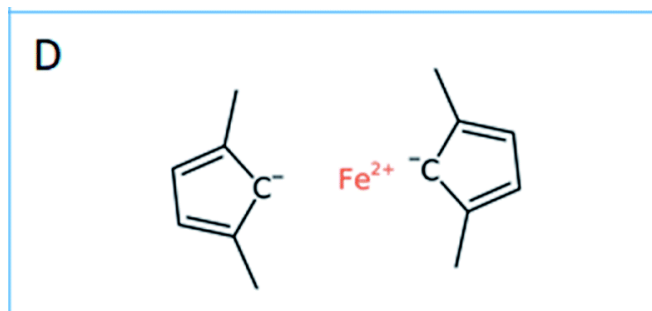
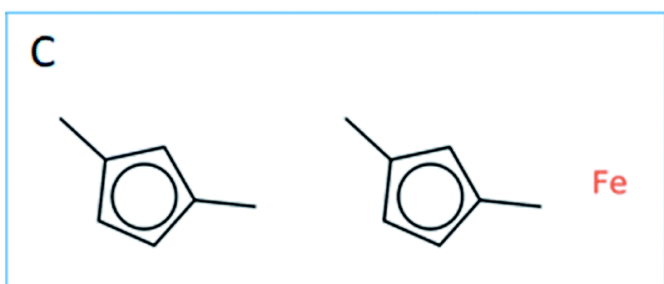
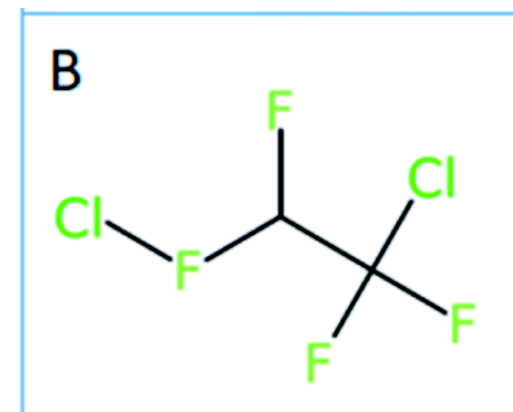
# PFAS in:	Definition A	Definition B	Definition C
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197

OntoChem PFAS lists	1,777,052
OntoChem PFAS from CORE - Definition A	26,805
OntoChem PFAS from CORE - Definition B	4,114
OntoChem PFAS from CORE - Definition C	3,432
OntoChem PFAS from Google Patents - Definition A	1,762,971
OntoChem PFAS from Google Patents - Definition B	73,749
OntoChem PFAS from Google Patents - Definition C	33,649



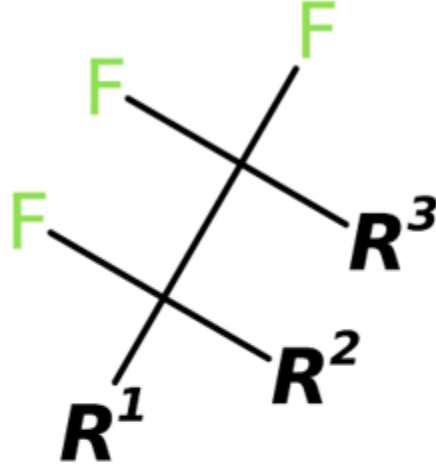
# Extracting PFAS (Def. B) from the 4.2 M Patent SMILES Set

Toolkit and normalizer selection (interim results)

Toolkit	Normalizer	PFAS	Not PFAS	Invalid
CDK	CDK Normalizer	78 412	4 104 264	36
OCL	OCL Standardizer	78 411	4 104 038	263
RDKit	molVS	75 762	3 988 584	118 366



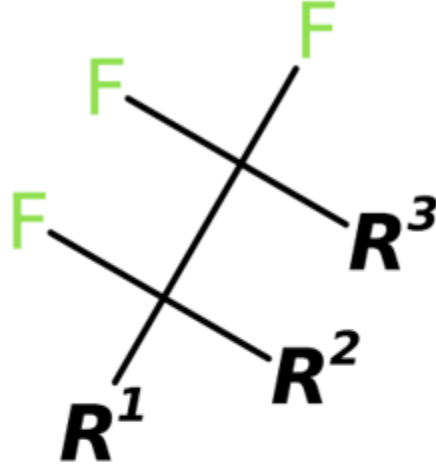


# Extracting PFAS – CORE and Patents, all definitions (OCL)

			
# PFAS in:	Definition A	Definition B	Definition C
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197



# Extracting PFAS – Complementarity of Data Sources

			
# PFAS in:	Definition A	Definition B	Definition C
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197
<b>Total</b>	<b>1,797,831</b>	<b>77,441</b>	<b>36,788</b>

# Comparing Extracted PFAS vs Existing PFAS Lists

OECD PFAS List (3.6K) *versus* CORE Definition A (27K)



- ▼ S25 | OECD PFAS | List of PFAS from the OECD ? 3,677
- ▼ Structure Category 3,677
  - ▶ 100 Perfluoroalkyl carbonyl compounds ( $C_nF_{2n+1}-C(=O)-R$ ) 490
  - ▶ 200 Perfluoroalkane sulfonyl compounds ( $C_nF_{2n+1}-S(=O)(=O)-R$ ) 458
  - ▶ 300 Perfluoroalkyl phosphate compounds ( $C_nF_{2n+1}-P(O)-R$ ) 16
  - ▶ 400 Fluorotelomer-related compounds 1,392
  - ▶ 500 Per- and polyfluoroalkyl ether-based compounds ( $C_nF_{2n+1}-O-C_mF_{2m+1}-R$ ) 322
  - ▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 282
  - ▶ 700 Other PFAA precursors or related compounds - semifluorinated 716
  - ▶ 800 Fluoropolymers 1

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>

Barnabas *et al.* (2022) *Digital Discovery*. DOI: [10.1039/D2DD00019A](https://doi.org/10.1039/D2DD00019A)



# Comparing Extracted PFAS vs Existing PFAS Lists

OECD PFAS List (3.6K) *versus* CORE Definition A (27K)



▼ S25 | OECDPFAS | List of PFAS from the OECD ? 3,677

▼ Structure Category 3,677

- ▶ 100 Perfluoroalkyl carbonyl compounds
- ▶ 200 Perfluoroalkane sulfonyl compounds
- ▶ 300 Perfluoroalkyl phosphate compounds
- ▶ 400 Fluorotelomer-related compounds
- ▶ 500 Per- and polyfluoroalkyl ether-based compounds
- ▶ 600 Other PFAA precursors and related compounds
- ▶ 700 Other PFAA precursors or related compounds
- ▶ 800 Fluoropolymers 1

▼ S25 | OECDPFAS | List of PFAS from the OECD ? 916

▼ Structure Category 916

- ▶ 100 Perfluoroalkyl carbonyl compounds (C<sub>n</sub>F<sub>2n+1</sub>-C(=O)-R) 120
- ▶ 200 Perfluoroalkane sulfonyl compounds (C<sub>n</sub>F<sub>2n+1</sub>-S(=O)(=O)-R) 181
- ▶ 300 Perfluoroalkyl phosphate compounds (C<sub>n</sub>F<sub>2n+1</sub>-P(O)-R) 7
- ▶ 400 Fluorotelomer-related compounds 344
- ▶ 500 Per- and polyfluoroalkyl ether-based compounds (C<sub>n</sub>F<sub>2n+1</sub>-O-C<sub>m</sub>F<sub>2m+1</sub>-R) 52
- ▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 129
- ▶ 700 Other PFAA precursors or related compounds - semifluorinated 83
- ▶ 800 Fluoropolymers

<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101>

Barnabas *et al.* (2022) *Digital Discovery*. DOI: [10.1039/D2DD00019A](https://doi.org/10.1039/D2DD00019A)





# Comparing Extracted PFAS vs Existing PFAS Lists

OECD PFAS List (3.6K) *versus* CORE Definition A (27K)

4:1  
ratio

▼ S25 | OECDPFAS | List of PFAS from the OECD ? 3,677

▶ 400 Fluorotelomer-related compounds 1,392

▶ 500 Per- and polyfluoroalkyl ether-based compounds ( $C_nF_{2n+1}-O-C_mF_{2m+1}-R$ ) 322

▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 282

▶ 700 Other PFAA precursors or related compounds - semifluorinated 716

▼ S25 | OECDPFAS | List of PFAS from the OECD ? 916

▶ 400 Fluorotelomer-related compounds 344

▶ 500 Per- and polyfluoroalkyl ether-based compounds ( $C_nF_{2n+1}-O-C_mF_{2m+1}-R$ ) 52

▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 129

▶ 700 Other PFAA precursors or related compounds - semifluorinated 83



# Comparing Extracted PFAS vs Existing PFAS Lists

OECD PFAS List (3.6K) *versus* CORE Definition A (27K)

4:1  
ratio

▼ S25 | OECDPFAS | List of PFAS from the OECD ? 3,677

▶ 400 Fluorotelomer-related compounds 1,392

▶ 500 Per- and polyfluoroalkyl ether-based compounds ( $C_nF_{2n+1}-O-C_mF_{2m+1}-R$ ) 322

▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 282

▶ 700 Other PFAA precursors or related compounds - semifluorinated 716

2:1  
ratio

▼ S25 | OECDPFAS | List of PFAS from the OECD ? 916

▶ 400 Fluorotelomer-related compounds 344

▶ 500 Per- and polyfluoroalkyl ether-based compounds ( $C_nF_{2n+1}-O-C_mF_{2m+1}-R$ ) 52

▶ 600 Other PFAA precursors and related compounds - perfluoroalkyl ones 129

▶ 700 Other PFAA precursors or related compounds - semifluorinated 83



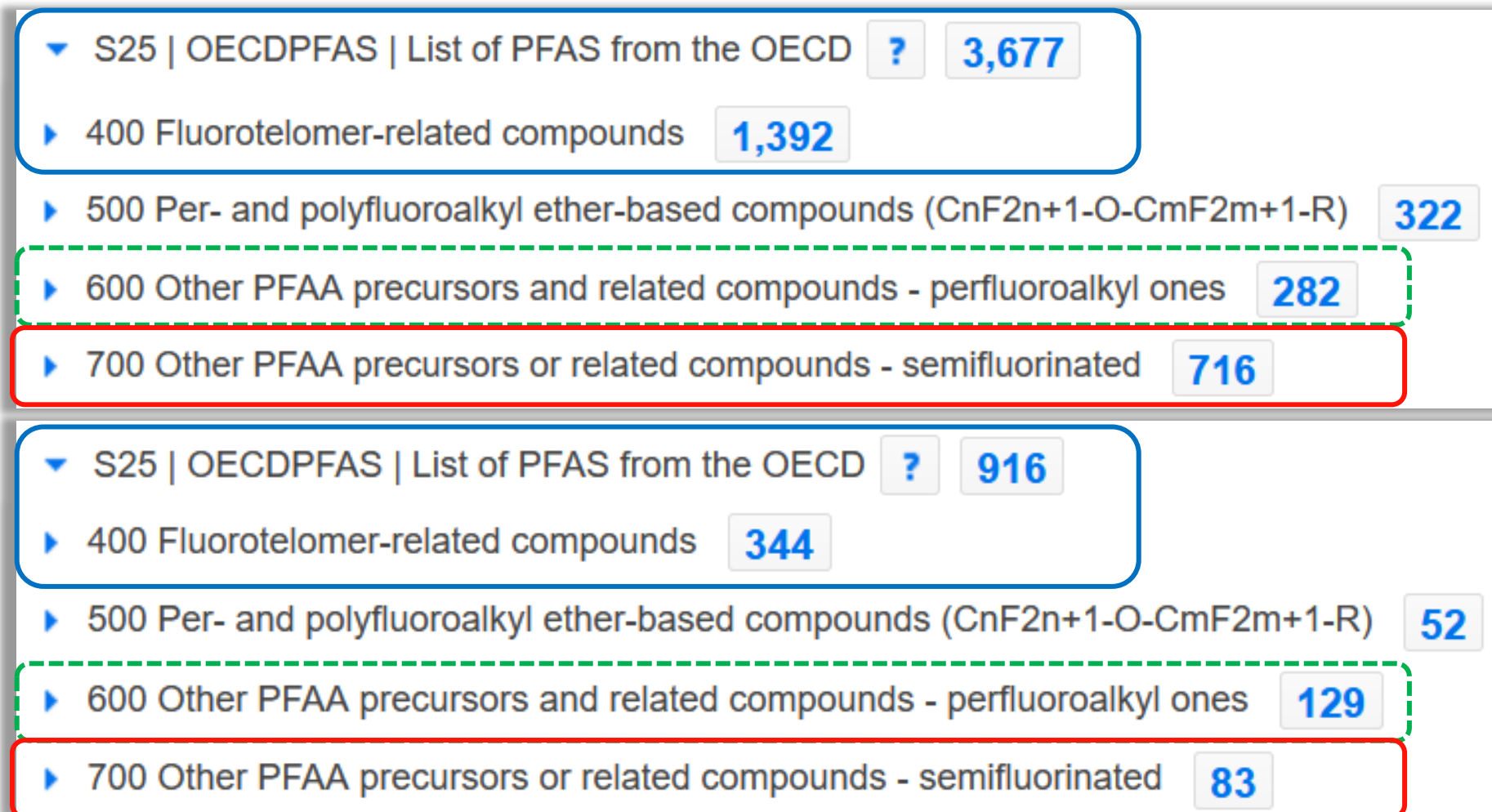
# Comparing Extracted PFAS vs Existing PFAS Lists

OECD PFAS List (3.6K) *versus* CORE Definition A (27K)

4:1  
ratio

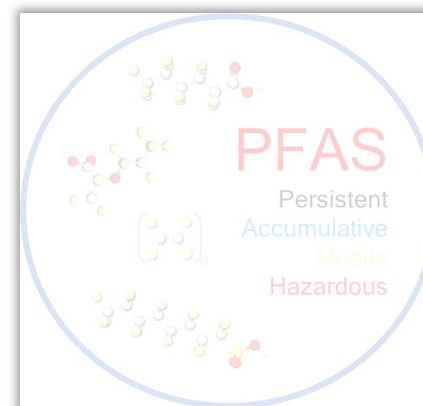
2:1  
ratio

~9:1  
ratio



# Outline for today

- Motivation
  - What are PFAS – and why?
- Extracting Chemical Data
- Different Toolkits and Definitions
- Helping People Find PFAS
- Questions / Discussion



	Definition A	Definition B	Definition C
# PFAS in:			
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197

▼ OntoChem PFAS lists [?](#) [↗](#) **1,777,052**

OntoChem PFAS from CORE - Definition A	<a href="#">?</a>	<a href="#">↗</a>	<b>26,805</b>
OntoChem PFAS from CORE - Definition B	<a href="#">?</a>	<a href="#">↗</a>	<b>4,114</b>
OntoChem PFAS from CORE - Definition C	<a href="#">?</a>	<a href="#">↗</a>	<b>3,432</b>
OntoChem PFAS from Google Patents - Definition A	<a href="#">?</a>	<a href="#">↗</a>	<b>1,762,971</b>
OntoChem PFAS from Google Patents - Definition B	<a href="#">?</a>	<a href="#">↗</a>	<b>73,749</b>
OntoChem PFAS from Google Patents - Definition C	<a href="#">?</a>	<a href="#">↗</a>	<b>33,649</b>

# Helping Researchers Find PFAS: MetFrag



MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

## Database Settings

Database: **OntoChem\_PFAS\_CORE**

Neutral Mass: 413.97365 Search ppm: 5

Formula: C<sub>8</sub>HF<sub>15</sub>O<sub>2</sub>

Identifiers:

Retrieve Candidates 17 Candidates

## Candidate Filter & Score Settings

## Fragmentation Settings & Processing

Mzppm: 5

Mzabs: 0.001

Mode: [M+H]

#	Molecule	Identifier	Mass	Formula	Normalized Scores	FinalScore	Details
1		OCID190000019728 InChIKeyBlock1 = SNGREZUHAYWORS	413.9737	C <sub>8</sub> HF <sub>15</sub> O <sub>2</sub>		5.9641	Peaks: 3 / 8 Fragments Scores Download
2		OCID190024173 InChIKeyBlock1 = FUGOJSZHPNU					
3		OCID190099936 InChIKeyBlock1 = KPBOPOXIOBB					

### Scores View

Candidate Identifier: OCID190000019728|53

	Name	Normalized Value	Raw Value
<input checked="" type="radio"/>	MetFrag	1.0	183.4503
<input checked="" type="radio"/>	ExactSpectralSimilarity	0.9641	0.9641
<input checked="" type="radio"/>	CORE_Refs	1.0	3417.0
<input type="radio"/>	Definition_A	1.0	1.0
<input type="radio"/>	Definition_B	1.0	1.0
<input type="radio"/>	Definition_C	1.0	1.0



# Helping Researchers Find PFAS: PubChem PFAS Tree



- PFAS and Fluorinated Compounds in PubChem [?](#) [↗](#) **19,692,732**
  - OECD PFAS definition [?](#) [↗](#) **6,203,598**
  - Organofluorine compounds [?](#) [↗](#) **19,543,301**
  - Other diverse fluorinated compounds [?](#) **114,526**
  - PFAS and fluorinated compound collections [?](#) [↗](#) **1,785,540**
    - CompTox Chemicals Dashboard PFAS suspect lists [?](#) [↗](#) **8,502**
    - NORMAN-SLE PFAS suspect lists [?](#) [↗](#) **6,040**
    - OntoChem PFAS lists [?](#) [↗](#) **1,777,052**
      - OntoChem PFAS from CORE - Definition A [?](#) [↗](#) **26,805**
      - PFAS from CORE - Definition B [?](#) [↗](#) **4,114**
      - PFAS from CORE - Definition C [?](#) [↗](#) **3,432**
      - OntoChem PFAS from Google Patents - Definition A [?](#) [↗](#) **1,762,971**
      - OntoChem PFAS from Google Patents - Definition B [?](#) [↗](#) **73,749**
      - OntoChem PFAS from Google Patents - Definition C [?](#) [↗](#) **33,649**
    - Other fluorinated chemical content in PubChem [?](#) [↗](#) **1,674**
  - NIST PFAS suspect list [?](#) [↗](#) **4,948**



Navigating over 6 million PFAS! A wa...

ZeroPM - H2020  
367 views •  
2 months ago



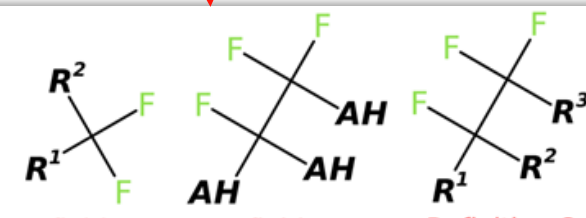
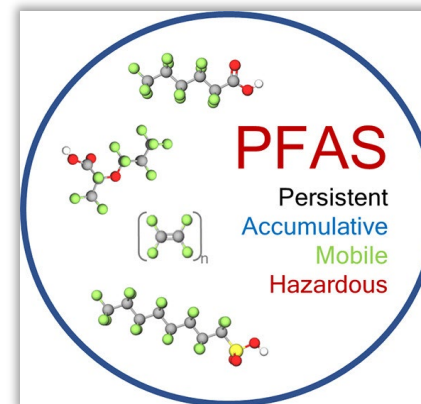
<https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120>

Barnabas et al. (2022) *Digital Discovery*. DOI: [10.1039/D2DD00019A](https://doi.org/10.1039/D2DD00019A)



# Outline for today

- Motivation
  - What are PFAS – and why?
- Extracting Chemical Data
- Different Toolkits and Definitions
- Helping People Find PFAS
- Questions / Discussion



	Definition A	Definition B	Definition C
# PFAS in:			
CORE Documents	27,958	4,139	3,457
Google Patents	1,783,651	75,108	34,197

▼ OntoChem PFAS lists [?](#) [↗](#) **1,777,052**

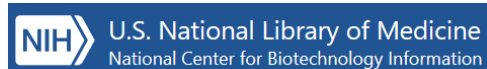
OntoChem PFAS from CORE - Definition A	<a href="#">?</a>	<a href="#">↗</a>	<b>26,805</b>
OntoChem PFAS from CORE - Definition B	<a href="#">?</a>	<a href="#">↗</a>	<b>4,114</b>
OntoChem PFAS from CORE - Definition C	<a href="#">?</a>	<a href="#">↗</a>	<b>3,432</b>
OntoChem PFAS from Google Patents - Definition A	<a href="#">?</a>	<a href="#">↗</a>	<b>1,762,971</b>
OntoChem PFAS from Google Patents - Definition B	<a href="#">?</a>	<a href="#">↗</a>	<b>73,749</b>
OntoChem PFAS from Google Patents - Definition C	<a href="#">?</a>	<a href="#">↗</a>	<b>33,649</b>

# Acknowledgements!

Email: [emma.schymanski@uni.lu](mailto:emma.schymanski@uni.lu)

Twitter: [@ESchymanski](https://twitter.com/ESchymanski)

Slides:



H2020: 101036756

