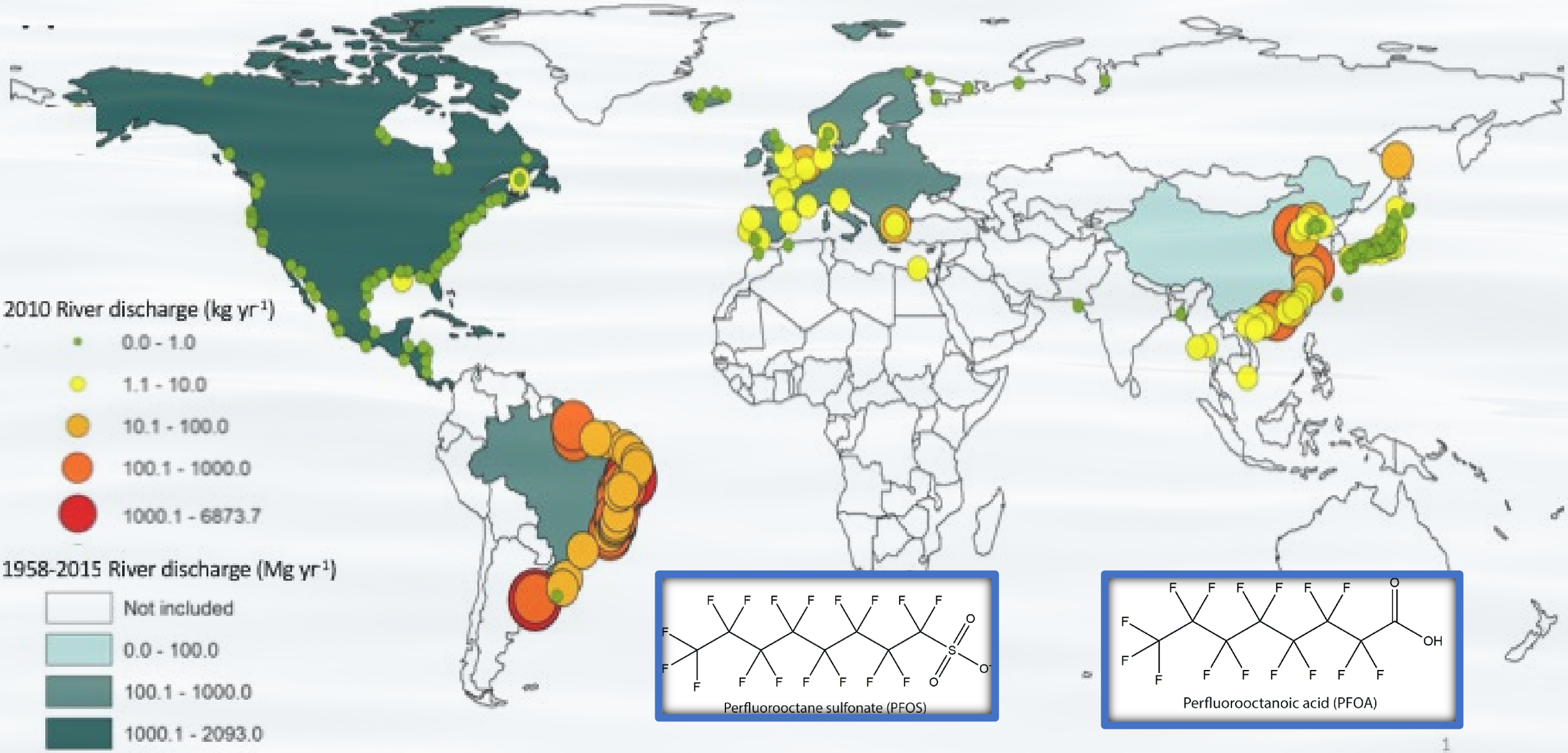
PFAS Toxprints:

A Hierarchical Structure-Based Categorization Method for Characterization of Per- and Polyfluoroalkyl Substances

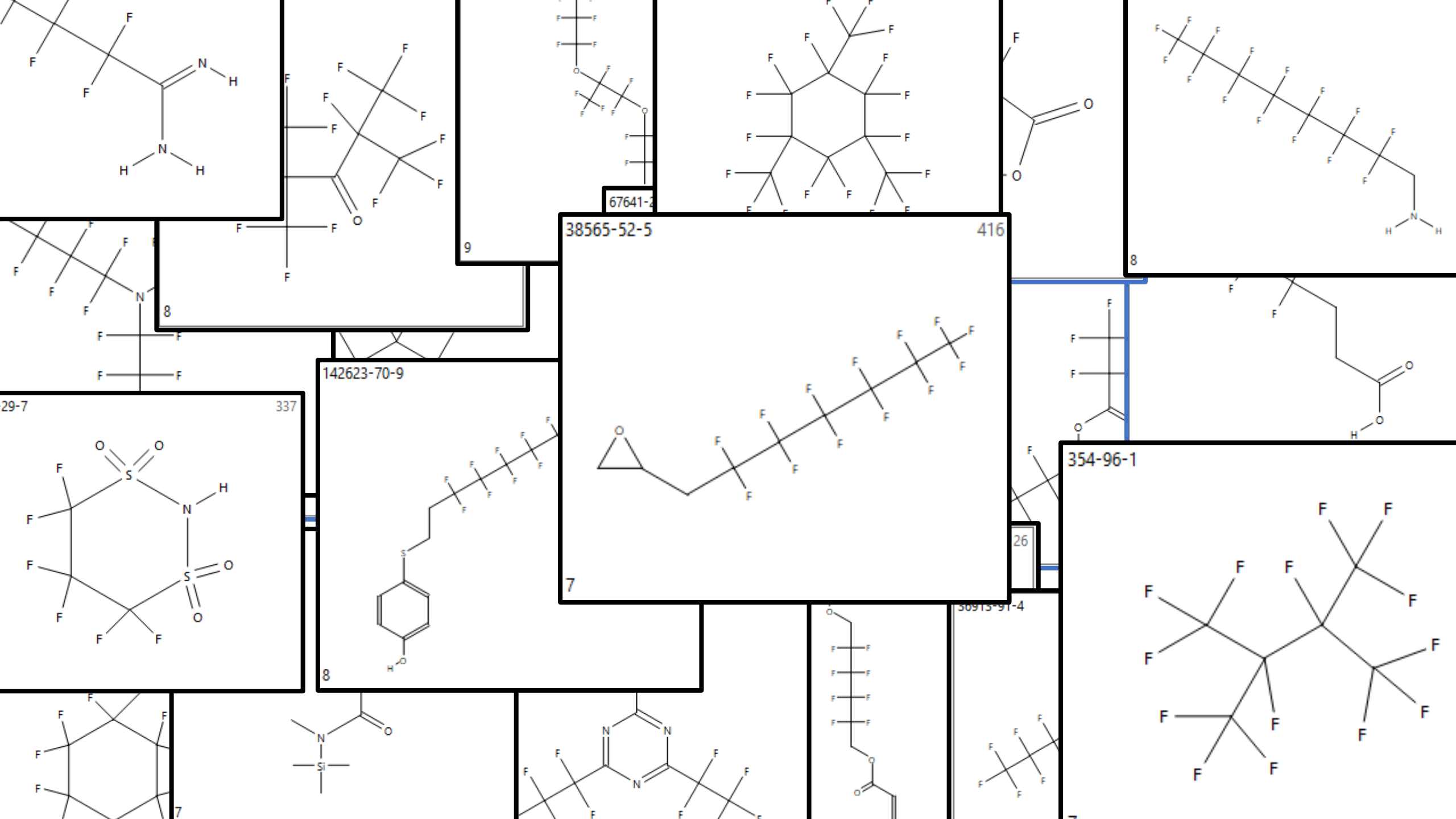
ORD/CCTE/CCCB
ORISE / Unied States Enviromnetal Protection Agency
Ryan Lougee

Oak Ridge Institute for Science and Education

Global PFOS river discharge



https://scholar.harvard.edu/files/styles/os_files_xlarge/public/ccwagner/files/global_pfos_riverdischarge_webpage.png?m=1549744187&itok=UITzT_qf


38565-52-5
416
7
142623-70-9
8
29-7
337
354-96-1
8
9
67641-2
26
36913-91-4
7

RCF2CFR'R" (R cannot be H)

PFAS|EPA: PFAS structures in DSSTox

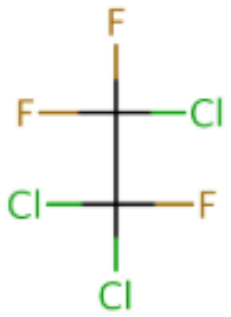
Search PFASSTRUCT Chemicals

Identifier substring search

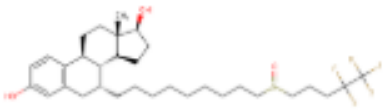
List Details

6648 chemicals

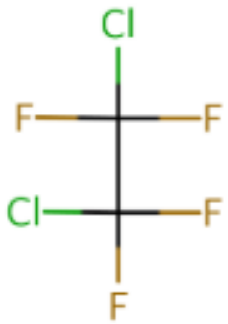
Select all | Download | Send to Batch Search | Default | DTXSID | CASRN | TOXCAST | Hide chemicals that are:



1,1,2-Trichloro-1,2,2-trifluoroethane
DTXSID:DTXSID6021377
CASRN:76-13-1

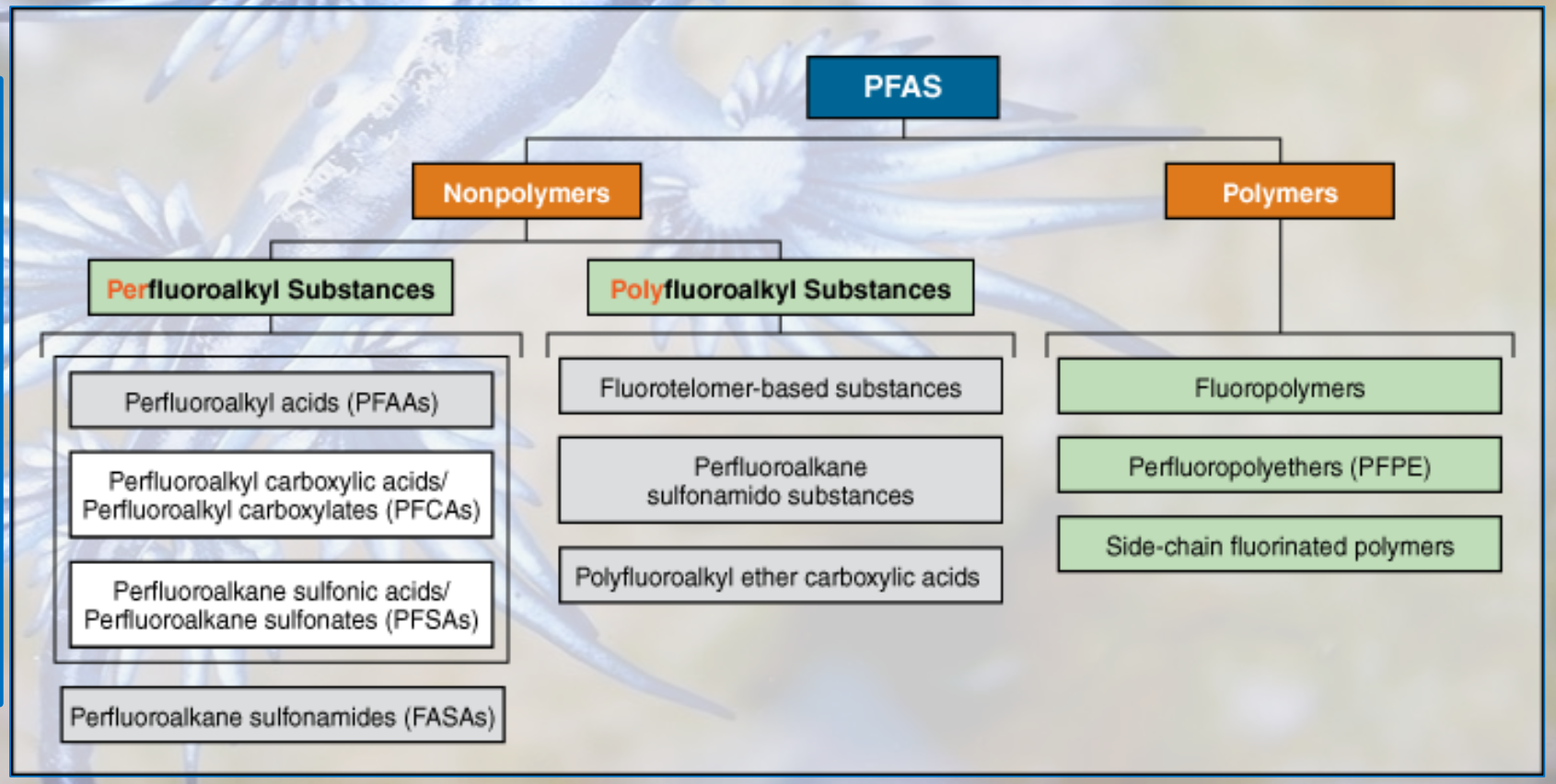


Fulvestrant
DTXSID:DTXSID4022369
CASRN:129453-61-8



1,2-Dichloro-1,1,2,2-tetrafluoroethane
DTXSID:DTXSID8026434
CASRN:76-14-2

How Can We Form a Greater Understanding of this Broad Chemical Category?



Why Categories?

- “Use Categories” could show how these effect the environment etc (surfactants)
- Specific groupings of structures seem to exhibit specific adverse effects C6-C8 chains for instance in literature
- The presence of certain functional groups sulfonyls and phosphates as well also effect adverse outcomes
- We can build categories for the breadth of byproducts, breakdown products, alternatives, and scaffold structures

Previous Attempts at Categorization

Integrated Environmental Assessment and Management — Volume 7, Number 4—pp. 513–541

Perfluoroalkyl and Polyfluoroalkyl Substances in the Environment: Terminology, Classification, and Origins

Robert C Buck,† James Franklin,*‡ Urs Berger,§ Jason M Conder,‖ Ian T Cousins,§ Pim de Voogt,#
Allan As
†E.I. du
‡CLF-Ch
§Depart
‖ENVIRO
#Institu
††Nordic
‡‡Wadsw
Health
§§Depar
‖‖RIKILT

(Submitte

ABST
The
the en
PFASs.
global

Environmental Science Processes & Impacts

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: Environ. Sci.: Processes Impacts, 2019, 21, 1835

Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs)†

Bo Sha,‡[a] Emma L. Schymanski,‡*[b] Christoph Ruttkies,[c] Ian T. Cousins[a]

Vol. 127, No. 1 | Brief Communication

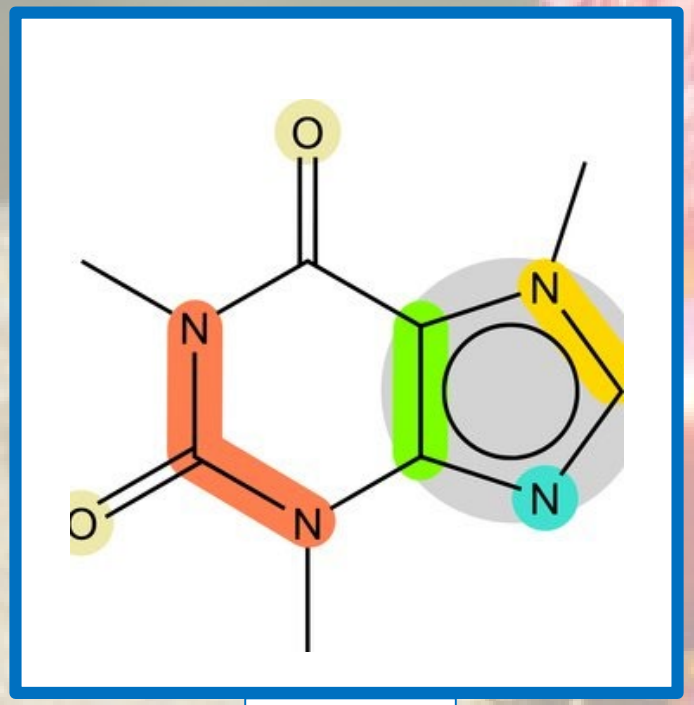
A Chemical Category-Based Prioritization Approach for Selecting 75 Per- and Polyfluoroalkyl Substances (PFAS) for Tiered Toxicity and Toxicokinetic Testing

Grace Patlewicz, Ann M. Richard, Antony J. Williams, Christopher M. Grulke, Reeder Sams, Jason Lambert, Pamela D. Noyes, Michael J. DeVito, Ronald N. Hines, Mark Strynar, Annette Guiseppi-Elie, and Russell S. Thomas

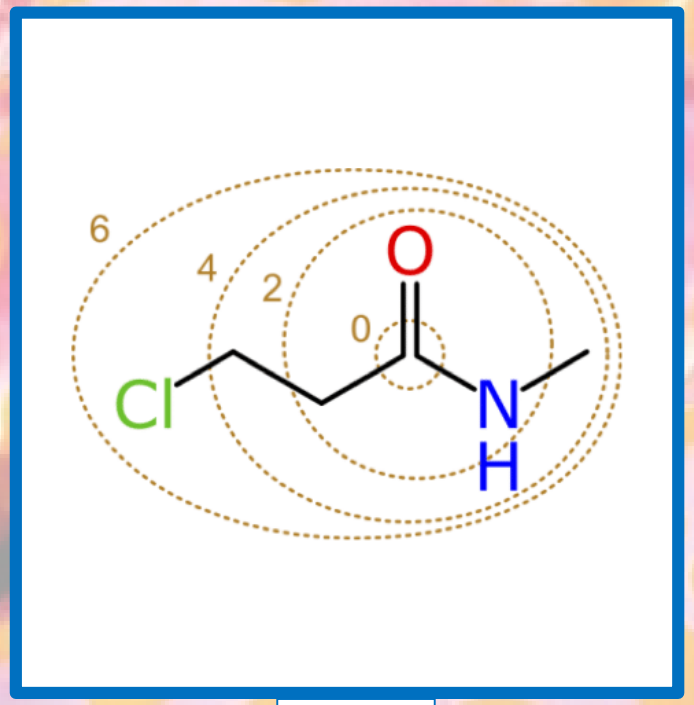
What Do We Need From Categories?

- Structure-Based
- Useful categories that reflect:
  - Adverse outcomes
  - Environmental outcomes
  - Byproducts
  - Others
- Reproducible
- Easy to use
- Enables Automation & Cheminformatics application

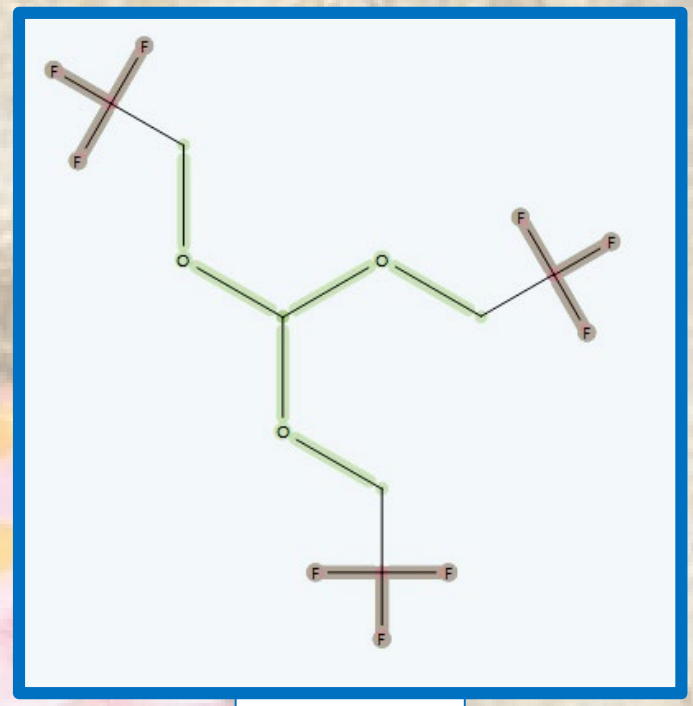
Molecular Fingerprints



MACCS



ECFP



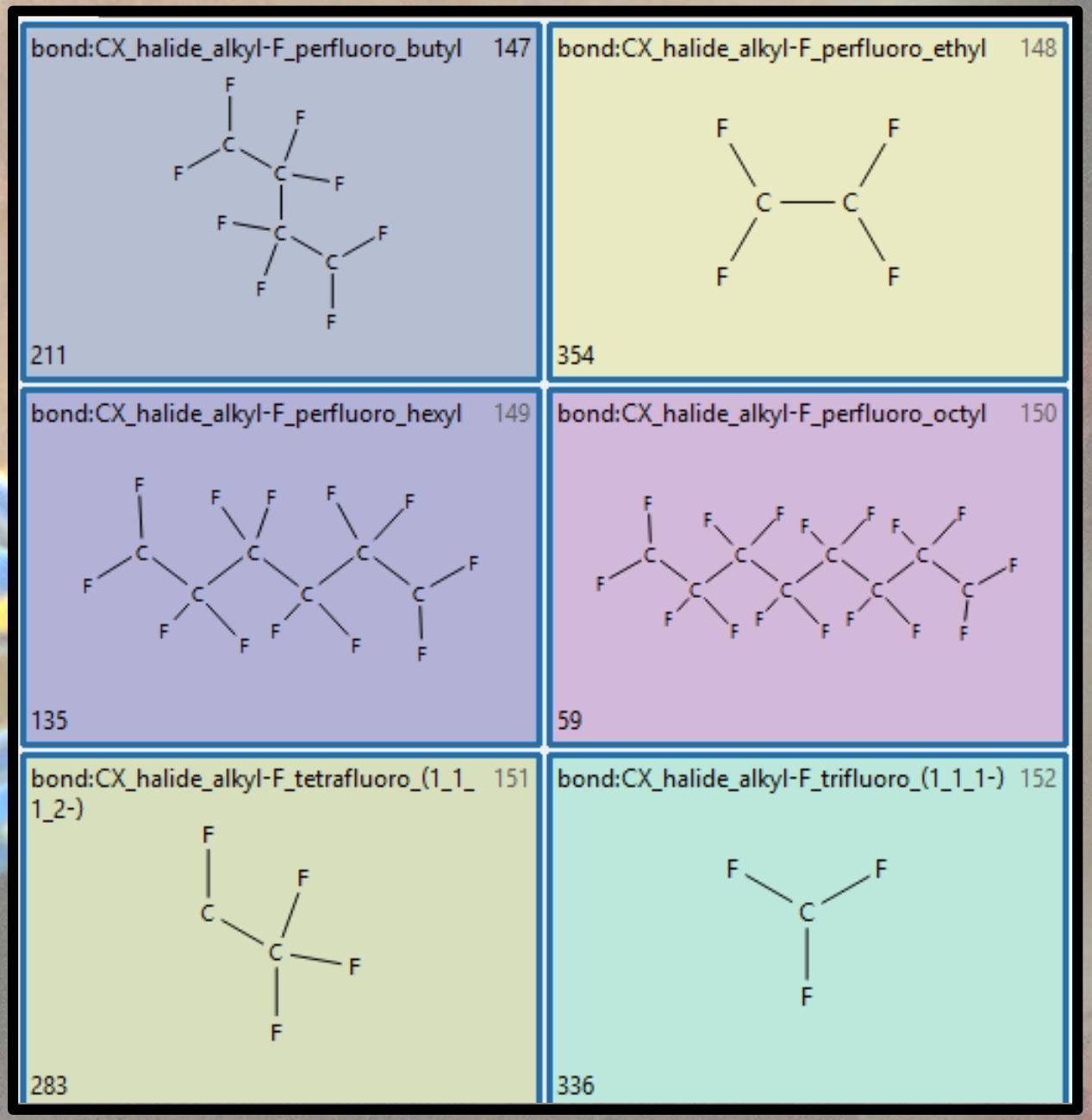
Toxprints

Output:

0100000101101110011001000111001001100101011101110010000001010111011010000110010101100101011011000110010101
0111001000100000011010010111001100100000011000010010000001100110011000010111010000100000011000100110100101
0111010001100011011010001

Toxprints

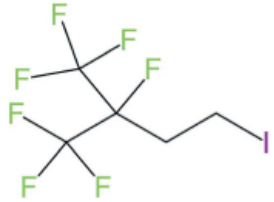
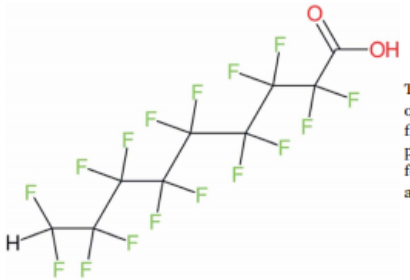
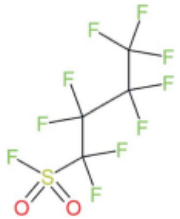
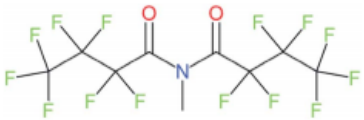
- 729 fragments
- PFAS substructures
- Good functional groups
- Some scaffolds



WHAT PFAS CONCEPTS ARE MISSING?

- FLUORINATED RINGS
- BRANCHING
- MULTIPLE R GROUPS
- POLYFLUORINATION NOT CAPTURED WELL
- ALTERNATIVE HALOGENATION
- MANY FUNCTIONAL GROUPS
- SPECIFIC CHAIN LENGTHS

Table 5 Selected cases outside the current scope of splitPFAS.

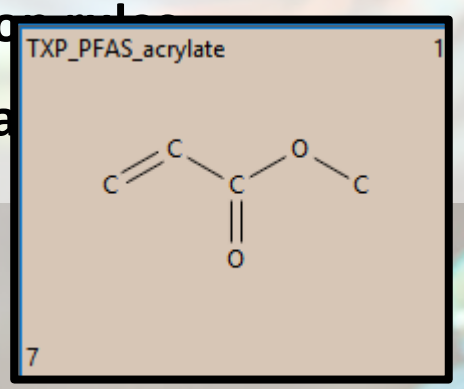
CAS_RN	Example structure	Explanation
Branched or cyclic perfluoroalkyl chains		
99324-96-6 Other examples: 28788-68-3 (ring)		This structure contains a branched perfluoroalkyl chain with two terminal CF_3 groups. To capture these, the default "pacs" SMARTS may need adjusting in future studies. It is likely that results for scenarios (iii) to (v) would be similar to those already observed
Polyfluoroalkyl (not perfluoroalkyl) chain		
76-21-1		The default "pacs" SMARTS in splitPFAS currently searches for C–C or C–F bonds, thus any structures with a non-C or F atom in the fluoroalkyl chain will not fulfil the pattern, like here where the pattern is H–(C_nF_{2n})–X–R, where here X = C(=O). Other members followed *e.g.* a Cl–(C_nF_{2n})–X–R pattern. These can be captured by adjusting the "pacs" option
The functional group R is F only		
375-72-4		These substances likewise failed the SMARTS pattern encoded into splitPFAS, which currently excludes compounds with a generic formula C_nF_{2n+1}–X–F. This could be addressed by adjusting the "pacs" option as well in future studies
Multiple R groups		
355-66-8		These examples were outside the scope defined for this article, examples of the form R_1–X–(C_nF_{2n})–X–R_2 are split correctly, but result in two PFAS chain results, which we did not consider further here
Multiple X Groups		
73980-71-9		For compounds in the form of (C_nF_{2n+1})X–R–X(C_mF_{2m+1}), the main issue is how to define C–X–R. There are built-in options to try various splitPFAS options in future studies

Time to Build Something New

Condition	Primary Level	Secondary Level	Tertiary Level
	TxP_PFAS_acrylate		
	bond:C(=O)O_carboxylicEster_alkenyl		
AND	bond:C(=O)O_carboxylicEster_acyclic		
AND	chain:alkeneLinear_mono-ene_ethylene_generic		
AND	bond:C=O_carbonyl_generic		
	TxP_PFAS_alcohol		
AND	bond:COH_alcohol_aliphatic_generic		
AND	bond:COH_alcohol_generic		
		TxP_PFAS_alcohol_primary	
AND		bond:COH_alcohol_pri-alkyl	
			TxP_PFAS_alcohol_primary_FT_diol
AND			bond:COH_alcohol_sec-alkyl
			TxP_PFAS_alcohol_primary_FTn1
AND			bond:C(~Z)~C~Q_a-haloalcohol
			TxP_PFAS_alcohol_primary_FTn2
AND			chain:alkaneLinear_propyl_C3
AND			chain:alkaneLinear_ethyl_C2(H_gt_1)
		TxP_PFAS_alcohol_polyF	
AND		bond:C(~Z)~C~Q_a-haloalcohol	
NOT		bond:COH_alcohol_pri-alkyl	
		TxP_PFAS_alcohol_sulfonylamide	
AND		bond:S(=O)N_sulfonylamide	

CSRML:

- Chemical Subgraphs and Reactions Markup Language (CSRML)
- XML based language
- Supports connectivity and topology but also properties of atoms, bonds, electronic systems
- Reactio
- Freewa



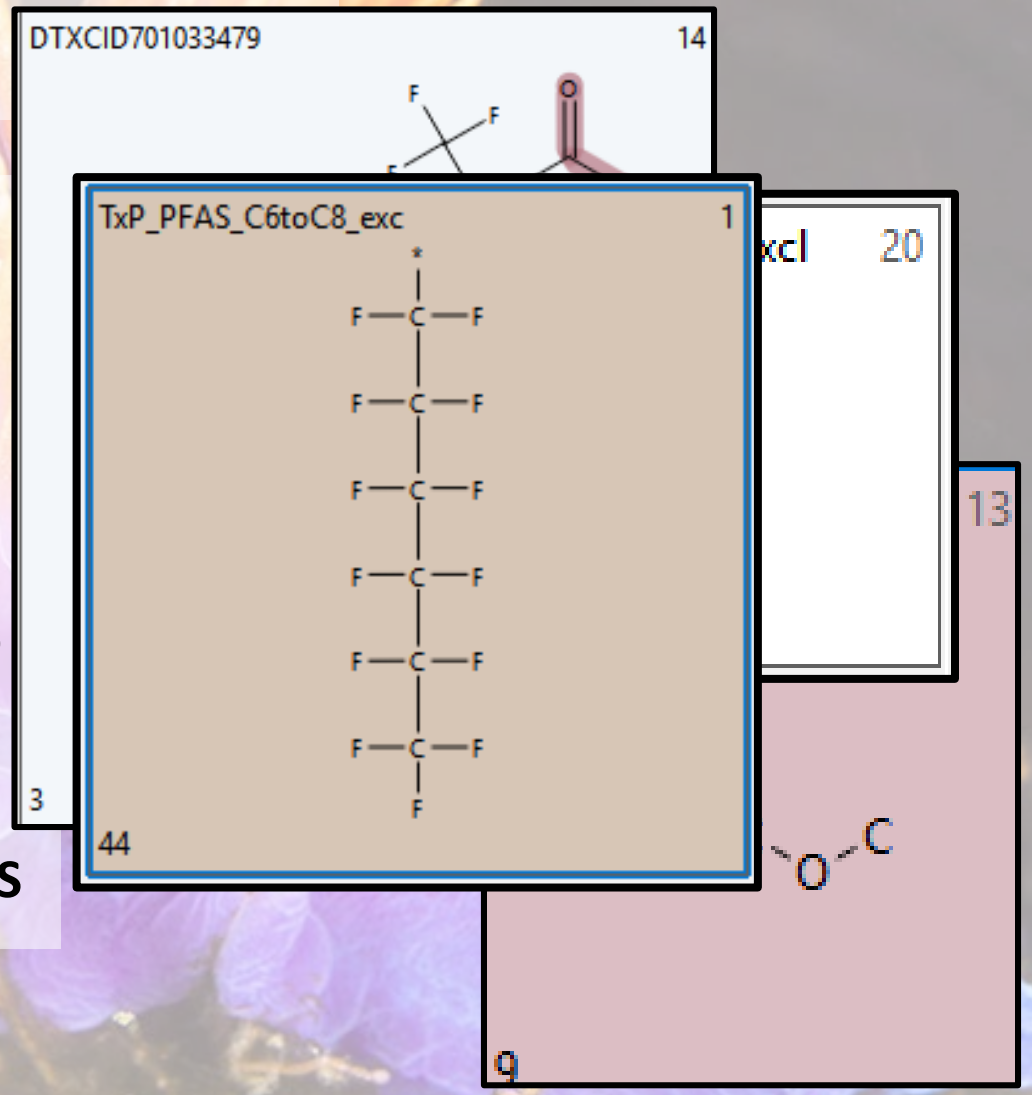

```
1005 > <classes id="ttc">…
1085   <subgraph id="TXP-000-0000-0000-0000-0000">
```

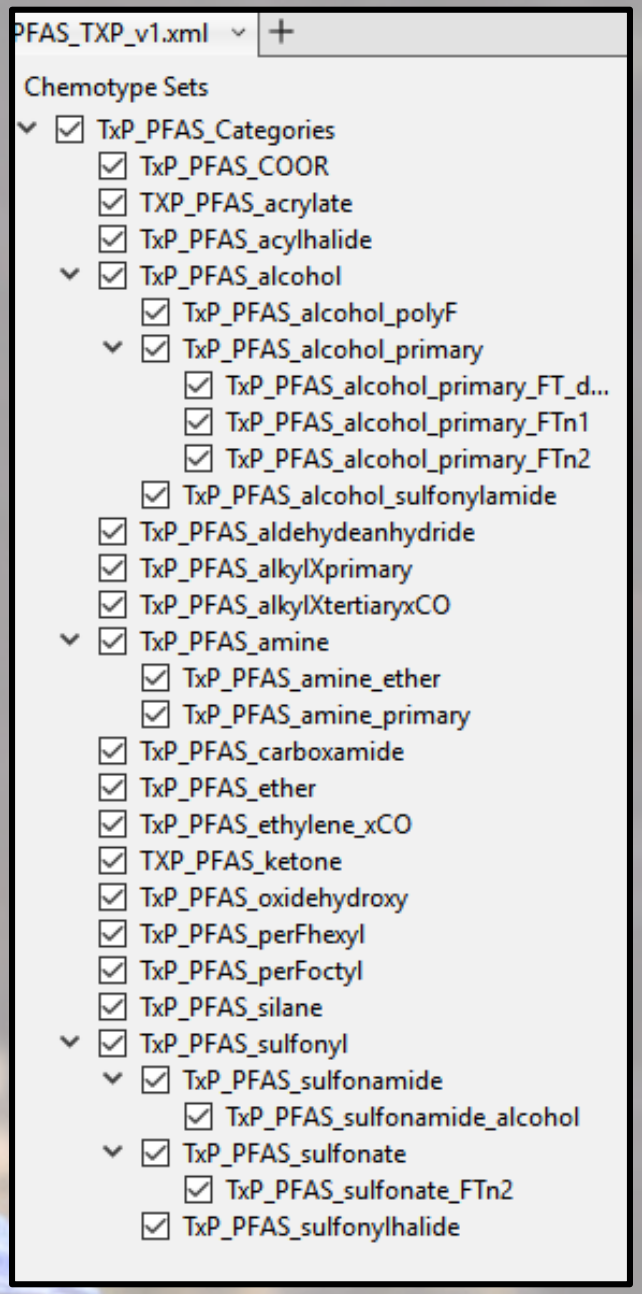
NEW_TXP_PFAS.xml | toxprint_V2.0_r711.xml | PFAS_TXP_v1.xml

```
1 <?xml version='1.0' encoding='utf-8'?>
2 <csrml xmlns="http://www.molecular-networks.com/schema/csrml" id="TxP_PFAS_Categories" csrmlVersion="2">
3   <title>
4     ToxPrint PFAS/PFOA Categories Version 1.0
5   </title>
6   <description>
7     $Id: PFAS_Categories.xml Lougee $
8     $Author: Lougee $
9   </description>
10
11 >  <classes id="TxP_PFAS_Categories">…
49    <!-- Acrylate from scratch not really just m179-->
50    <subgraph id="TxP_PFAS_acrylate">
51      <label>TXP_PFAS_acrylate</label>
52      <title>TXP_PFAS_acrylate</title>
53      <comment>TXP_PFAS_acrylate</comment>
54      <molecule id="m188">
55        <matchIf feature="substructureMatch"/>
56        <atoms>
57          <atom element="C" x="1.7411" y="0.9999" id="a1"/>
58          <atom element="O" x="2.6203" y="1.4998" id="a2"/>
59          <atom element="O" x="1.7411" y="0.0" id="a3"/>
60          <atom element="C" x="3.4823" y="0.9999" id="a4"/>
61          <atom element="C" x="0.8792" y="1.4998" id="a5"/>
62          <atom element="C" x="0.0" y="0.9999" id="a6"/>
63        </atoms>
64        <bonds>
65          <bond order="single" id="b1">
66            <atom id="a1"/>
67            <atom id="a2"/>
68          </bond>
69          <bond order="double" id="b2">
70            <atom id="a1"/>
```

INTERESTING THINGS ABOUT CSRML:

- **HIERARCHY**
- **MULTIPLE DISTINCT SUB-STRUCTURES**
- **INTERESTING ATOM AND BOND TYPES**
  Ex: Ring & Chain Atom
- **CUSTOMIZABLE ATOM AND BOND TYPES**
- **SPECIFIC CHAIN LENGTHS**
- **RANGES OF CHAIN LENGTHS**





How These Were Built: Structure Aggregation

- Searching Through literature to find interesting byproducts and structures
- Structures related to Adverse Outcomes
- Buck et al expert categories
- OECD category structures
- Missing OECD categories
- Once these were built I could filter out functional groups and structural groups and see what may still be missing
- Finally, generalized groups were added to capture broader categories

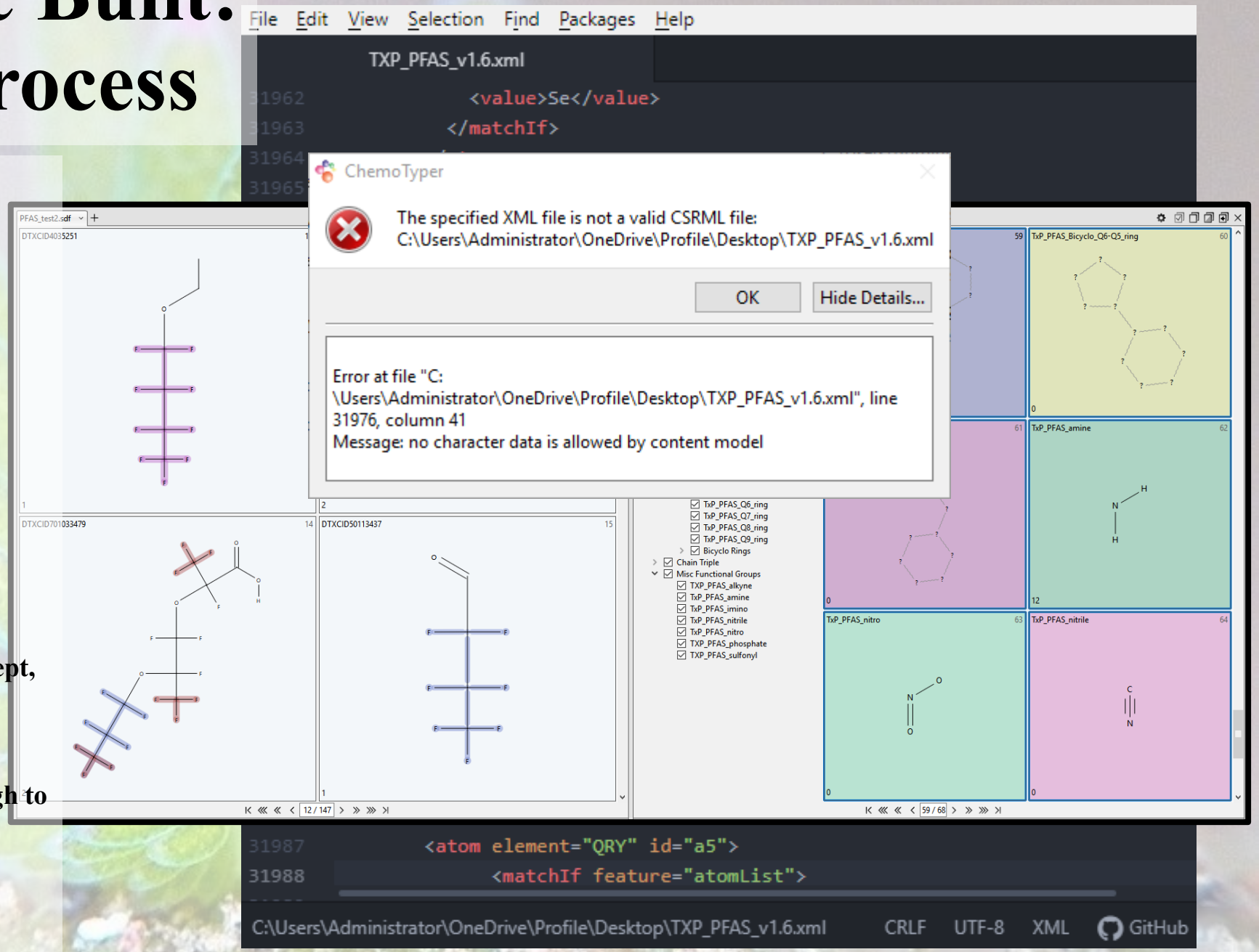
How Were These Built: Programming Process

- Syntax similar to XML
- Loaded into an IDE
- Looked for similar structure to what I was interested in
- Examined code
- Repurposed it
- Tested it in the Chemotyper
- Resolved loading errors

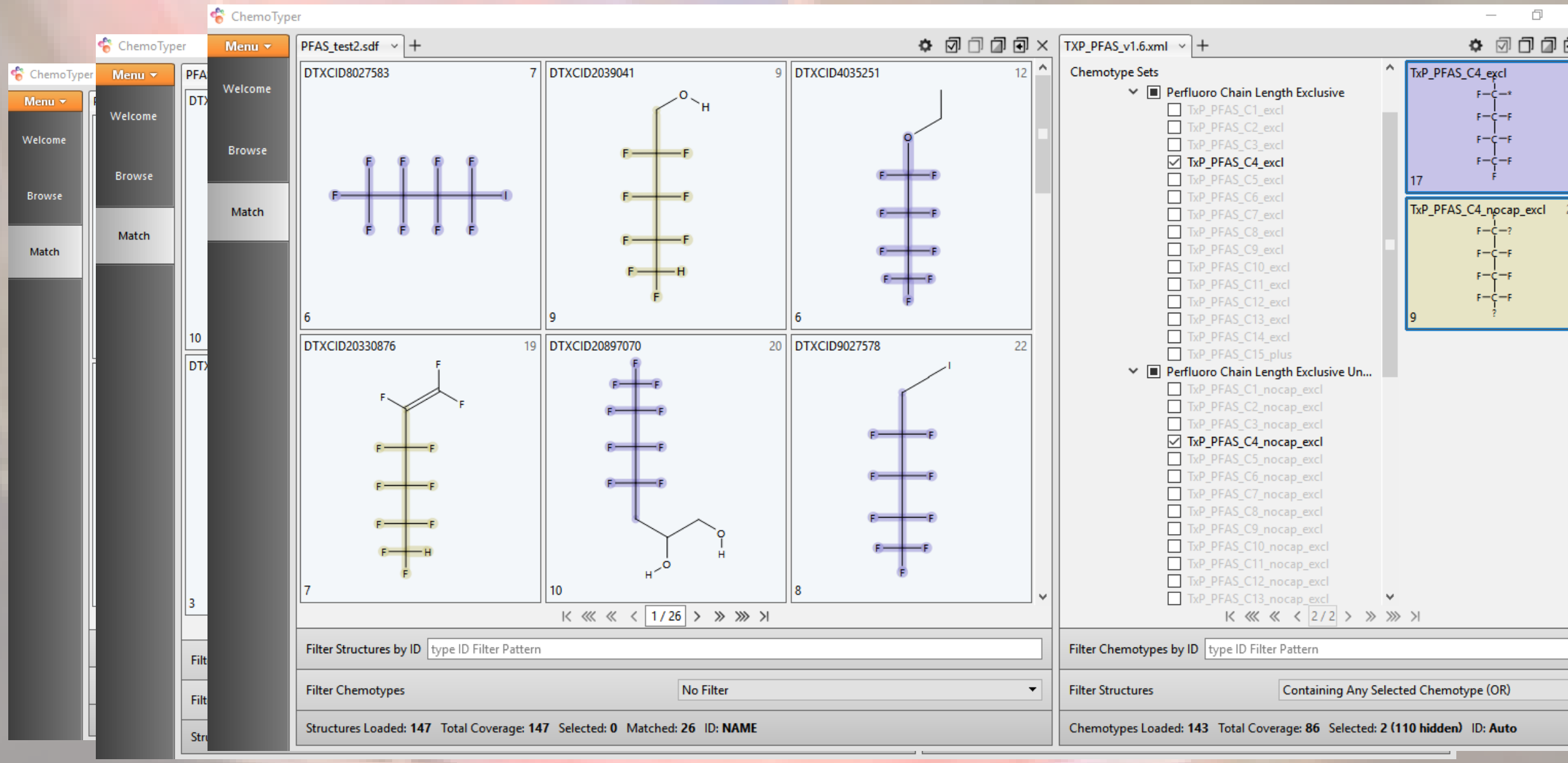
- Once structures loaded correctly, checked against dataset of PFAS to see that they correctly captured intended chemical concept, and chemotype looked correct

- Eventually, understood CSRML well enough to construct new concepts

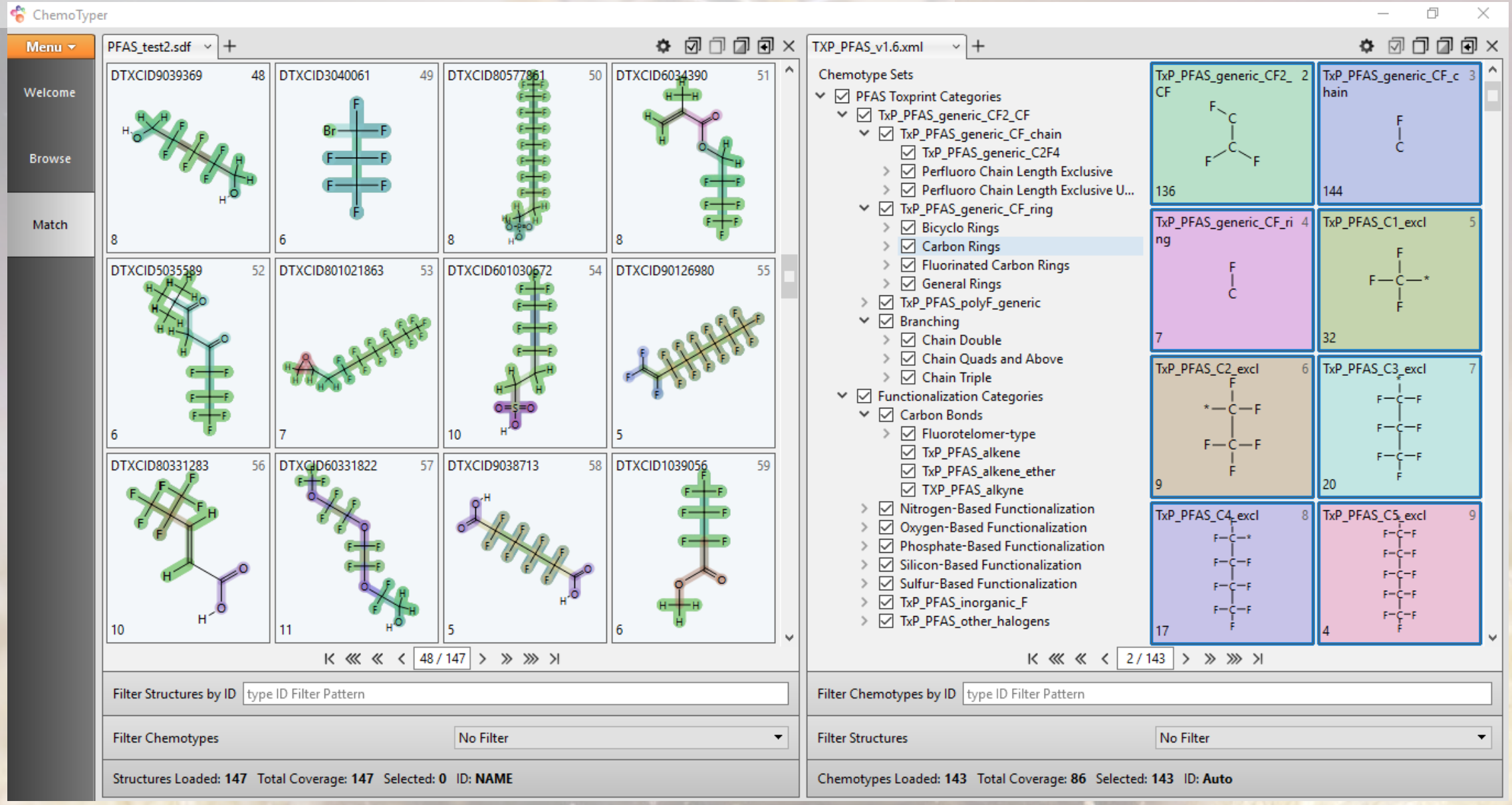
- Lastly, encoded the hierarchy



Some Things I Like



How to use these now?



How to use these now?

Register | Login

Wiki

Eventually here: https://toxprint.org/#ToxPrintChemotypes

Advanced Search

Backlogs

Bug Backlog
Feature Requests

Wiki

Wiki Homepage

ToxPrint Chemotypes

Acknowledgement

The ChemoTyper organizes the current version ToxPrint chemotypes into three functional areas:

1. Generic Structural Fragments
2. Structural Rules and Alerts
3. Category Classifiers

Generic Structural Fragments

Generic structural fragments are organized by atom, bond, chain, ring types as well as chemical groups including amino acids, carbohydrates, ligands, and nucleobases based on 729 essential chemotypes of the current ToxPrint_v2.0_r1520.xml (whatever the file name). These chemotypes can be generated as chemical fingerprints, either in binary (0/1) or counts data. They can be used to calculate similarity measures or structural feature descriptors for building models. (Yang 2015)

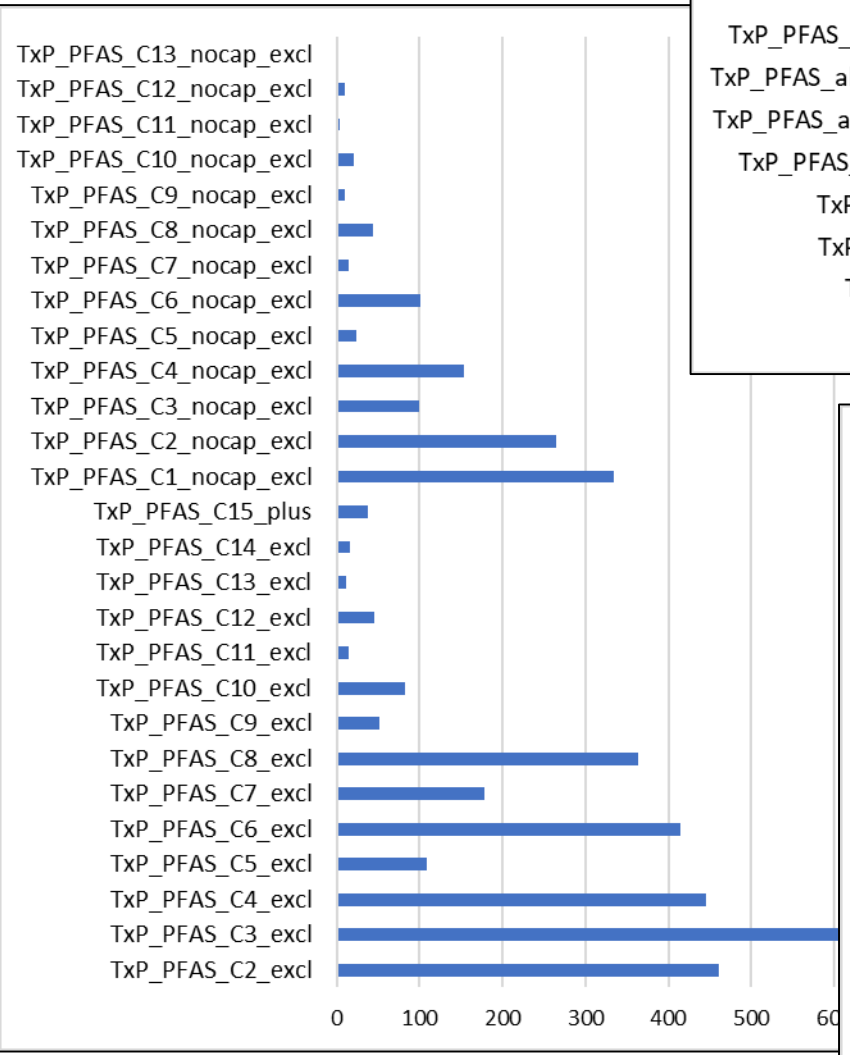
Structural Rules and Alerts

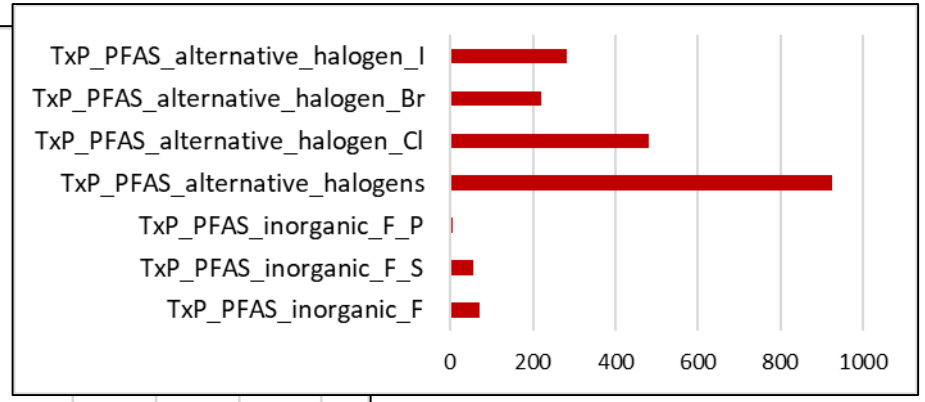
These can be developed using ToxPrint chemotypes as building blocks. The chemotypes defined in the ToxPrint set can be further refined or coded with properties (atom, bond, molecular, or physicochemical) to constrain the matches in order to enhance the signal-to-noise ratio of ToxPrint chemotypes when profiling the biological observations. To this end, we are developing ChemoType Editor to empower the users with the ability to fluently manipulate the CSRML query definitions graphically in a molecular editor. Please contact MN-AM if you are interested.

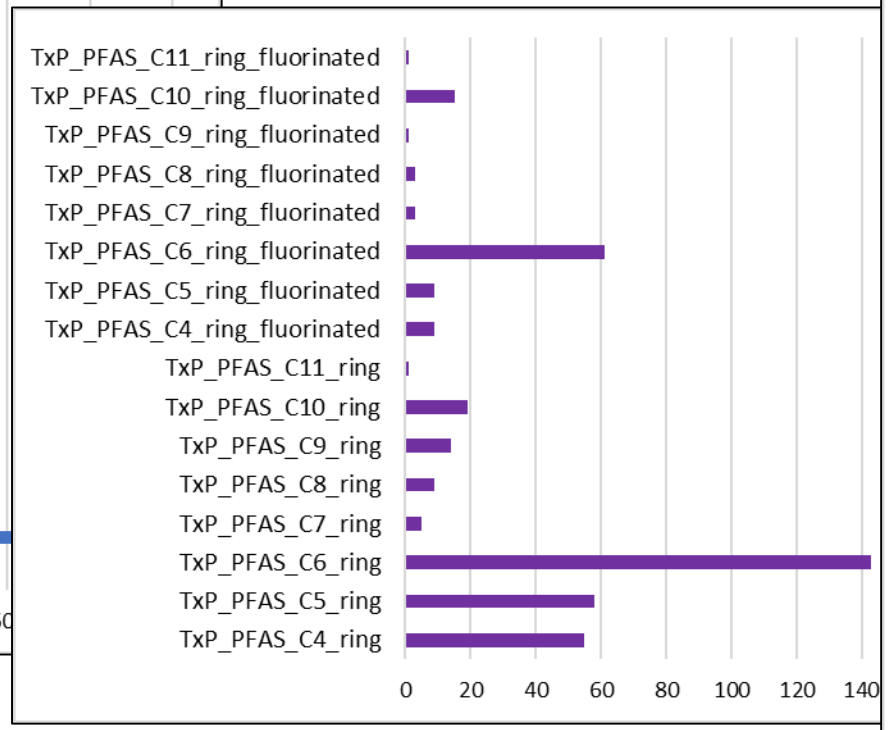
- Ashby-Tennant Genotoxic Carcinogen Alerts
- DNA binders
- Protein binders
- General Liver Alerts

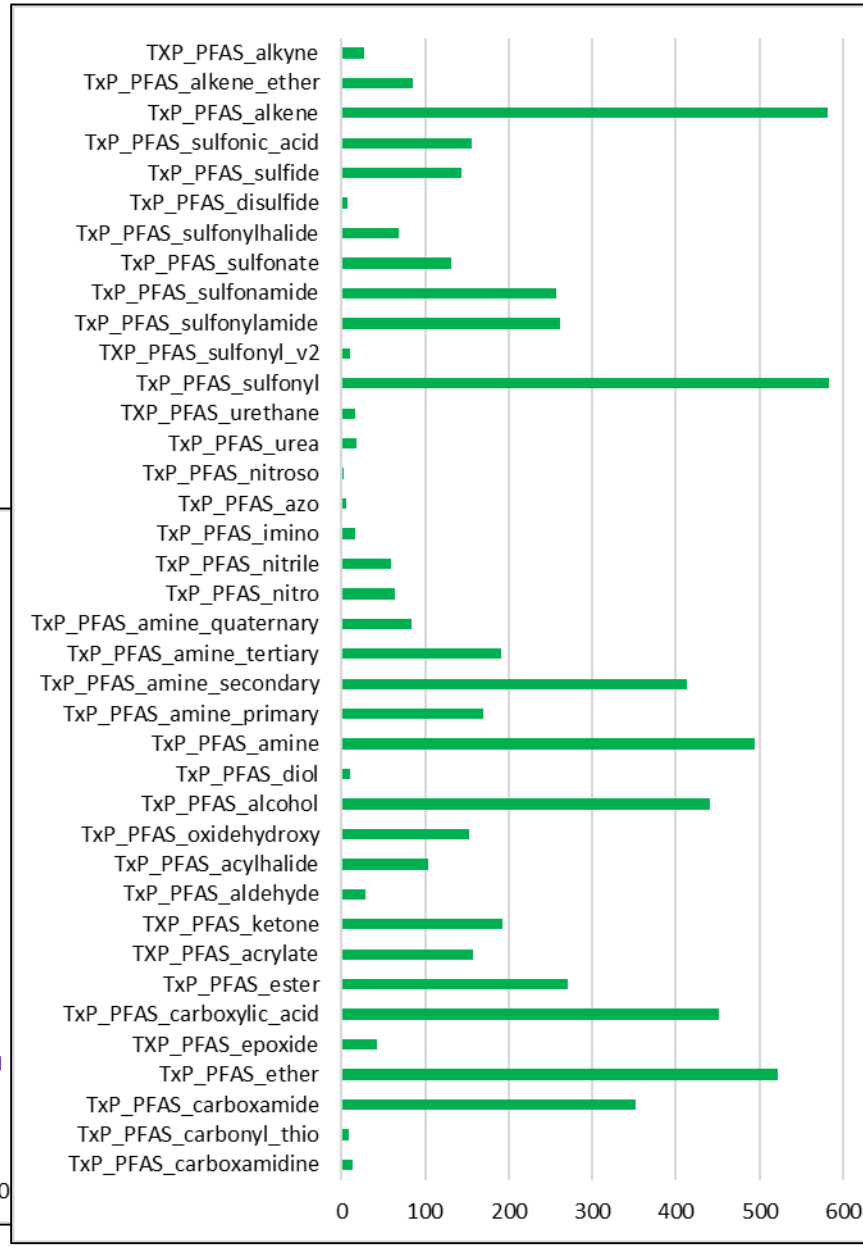
Lougee.Ryan@epa.gov

OECD PFAS PROFILE









THANK YOU