

PFAS Toxprints:

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

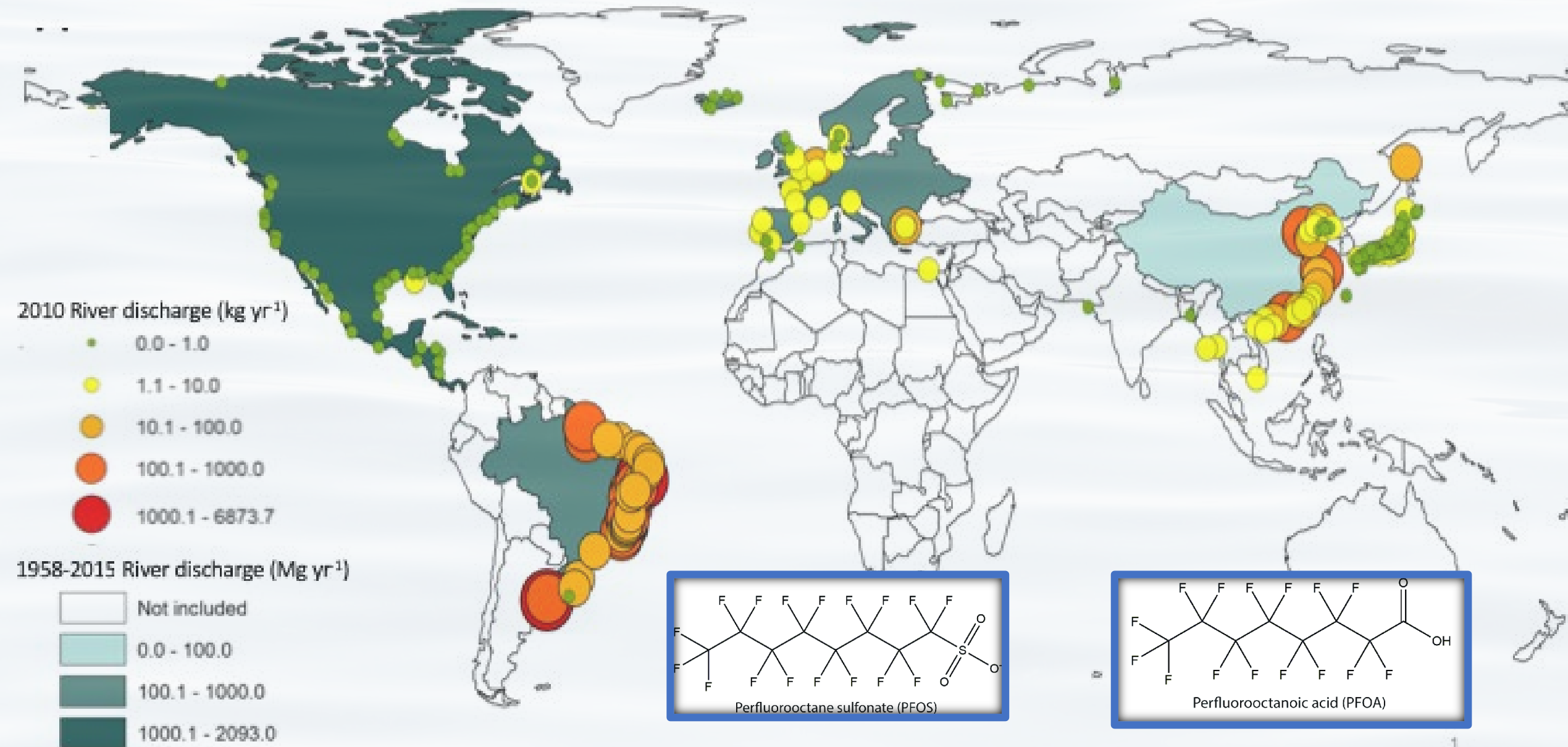
ORD/CCTE/CCCB
ORISE / United States Environmental Protection Agency
Ryan Lougee

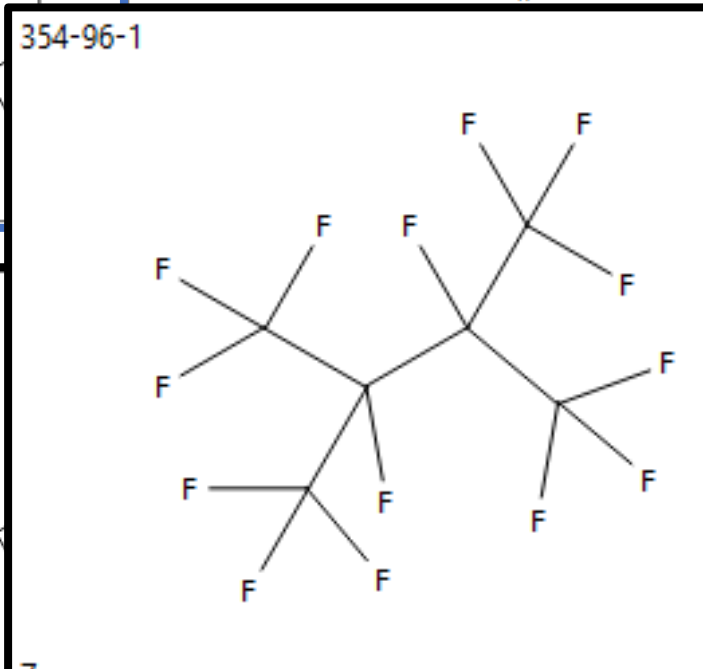
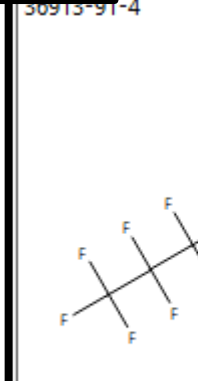
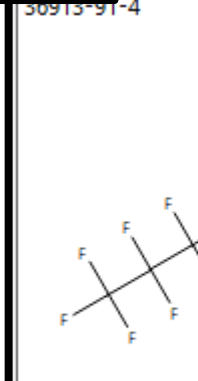
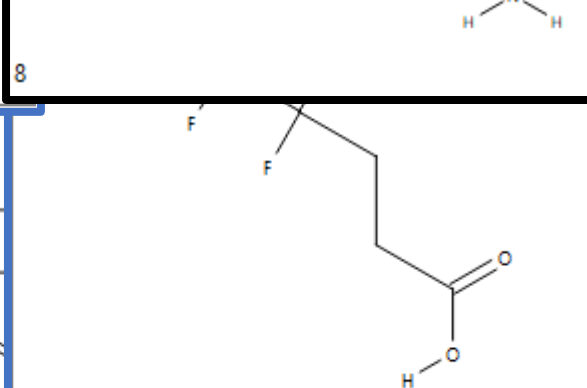
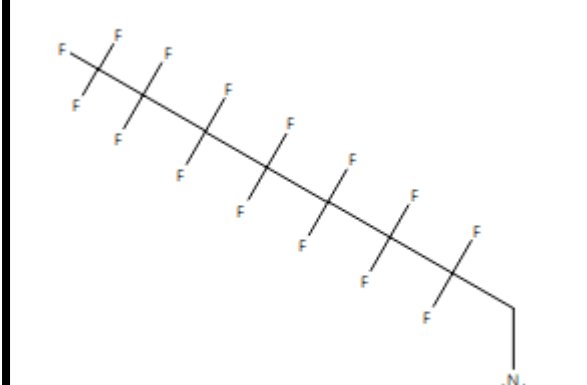


Oak Ridge Institute for Science
and Education

The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the U.S. EPA

Global PFOS river discharge





RCF2CFR'R" (R cannot be H)

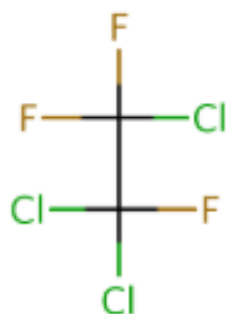
PFAS|EPA: PFAS structures in DSSTox

☐ 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

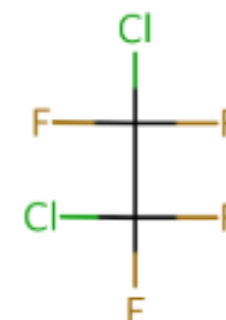
CASRN:76-13-1



Fulvestrant

DTXSID:DTXSID4022369

CASRN:120453-61-8

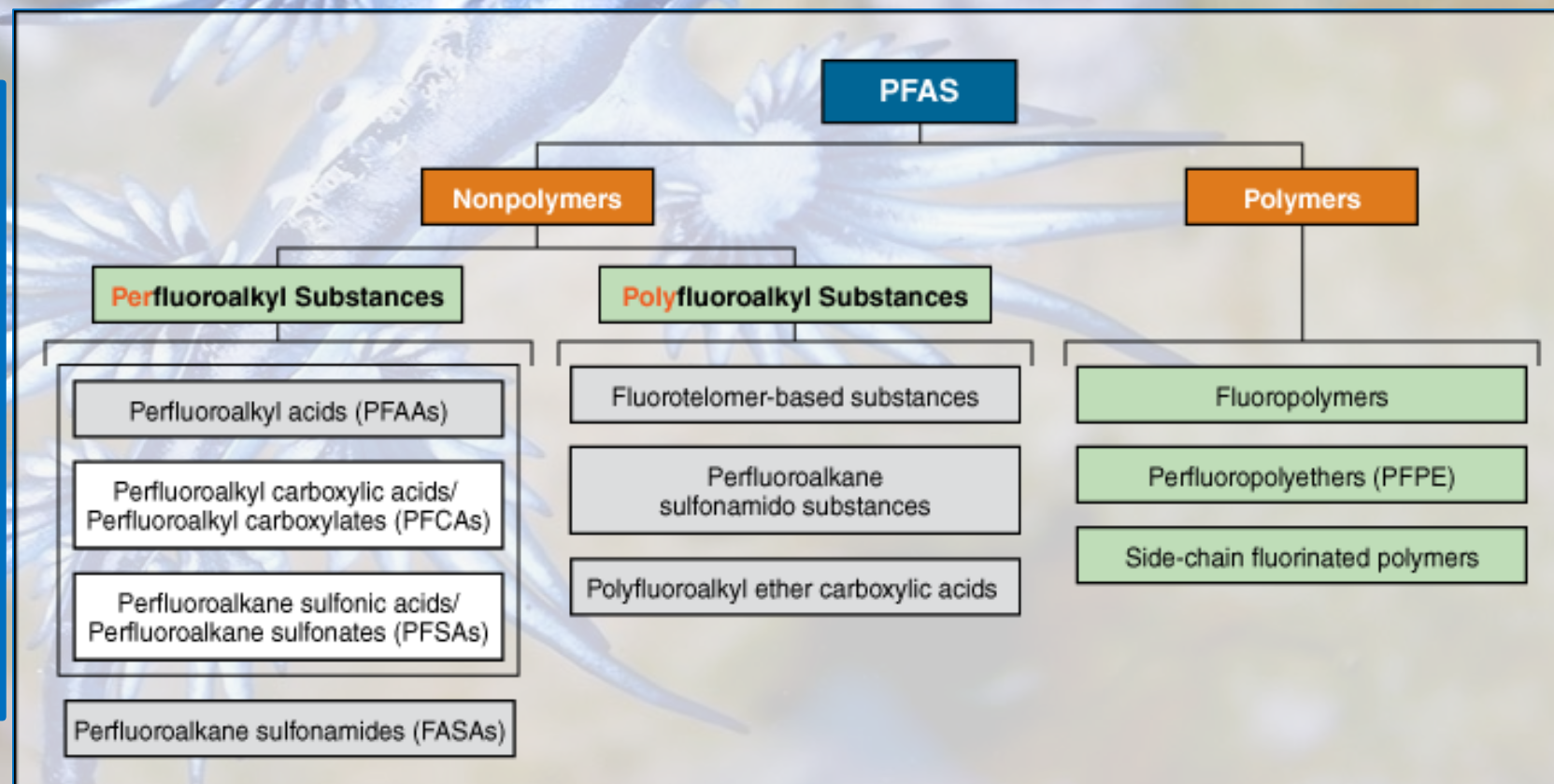
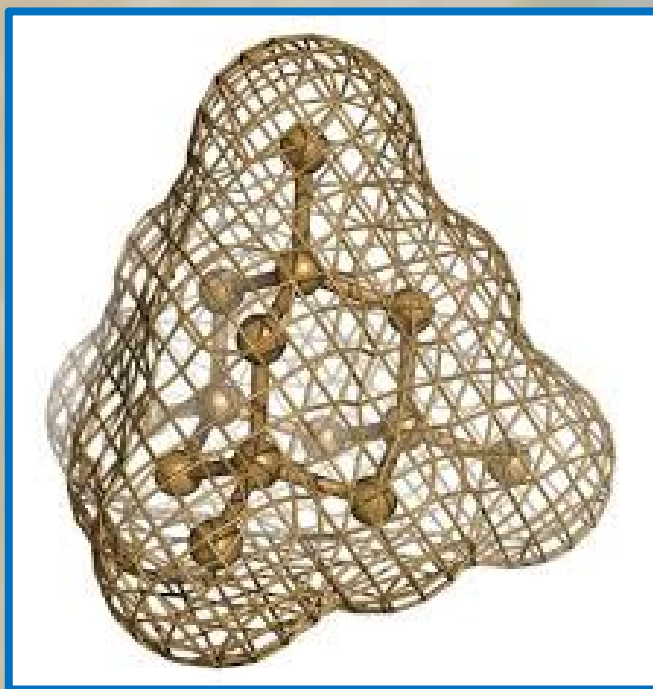


1,2-Dichloro-1,1,2,2-tetrafluoroethane

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
© 2011 SETAC

513

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 Vooqt,# Allan As

†E.I. du
‡CLF-Ch
§Depart
||ENVIR
#Institu
††Nordic
‡‡Wadsw
Health
§§Depart
|||RIKILT

(Submitted

ABST

The
the en
PFASs.
global

19, Downloaded on 1/7/2020 2:23:18 PM.
e Commons Attribution 3.0 Unported Licence.

Environmental Science Processes & Impacts



PAPER

[View Article Online](#)
[View Journal](#) | [View Issue](#)



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

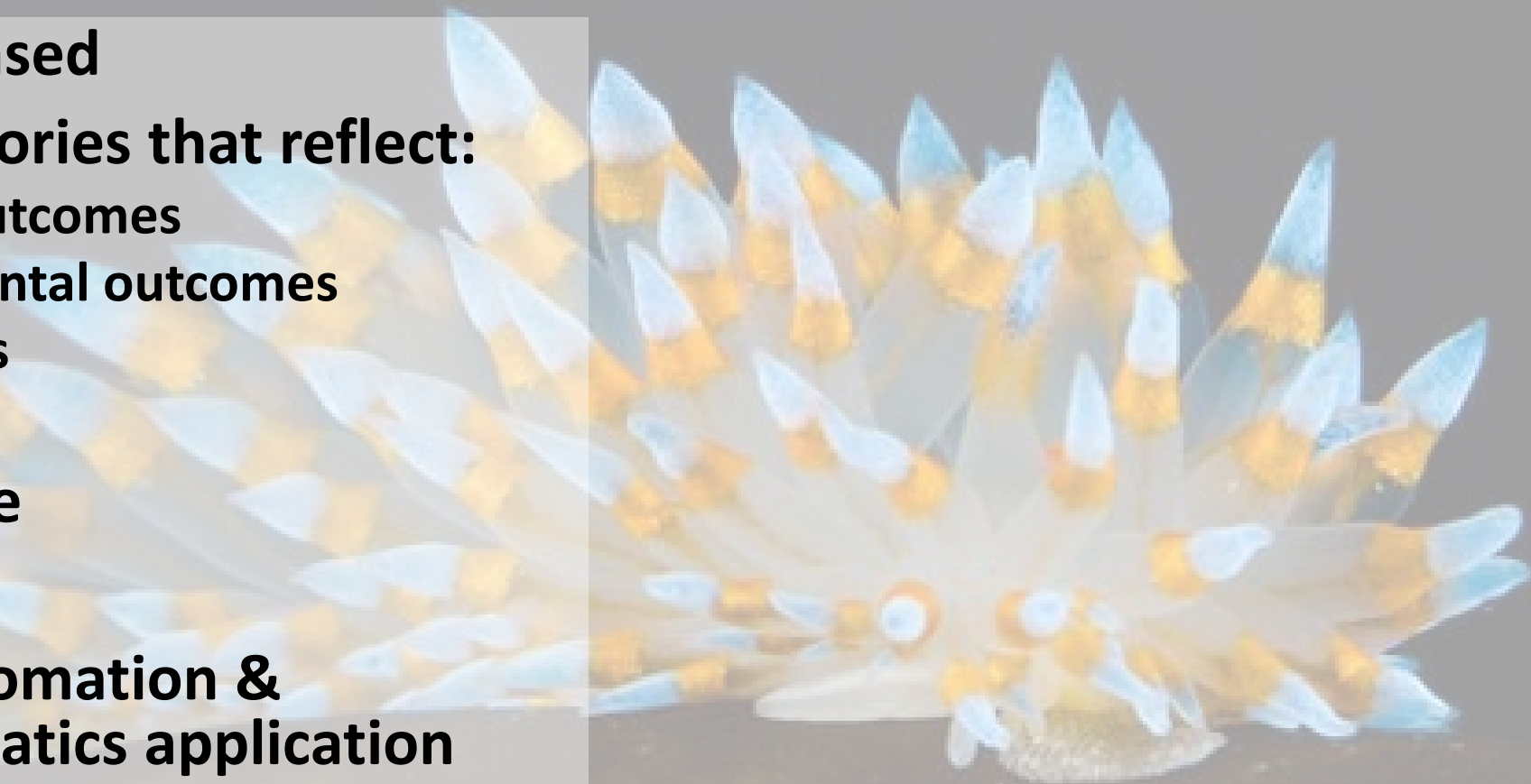
Vol. 127, No. 1 | Brief Communication

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

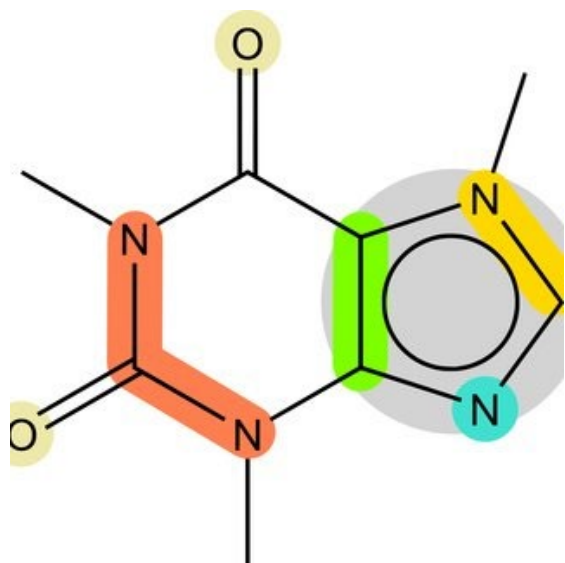
Grace Patlewicz , Ann M. Richard, Antony J. Williams, Christopher M. Grulke, Reeder Sams, Jason Lambert, Pamela D. Noyes, Michael J. DeVito, Ronald N. Hines, Mark Strynar, Annette Guiseppi-Elie, and Russell S. Thomas

What Do We Need From Categories?

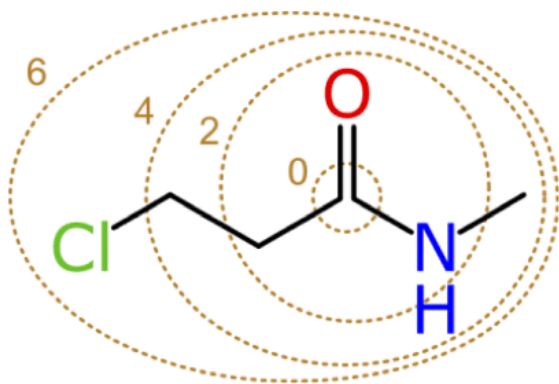
- **Structure-Based**
- **Useful categories that reflect:**
 - Adverse outcomes
 - Environmental outcomes
 - Byproducts
 - Others
- **Reproducible**
- **Easy to use**
- **Enables Automation & Cheminformatics application**



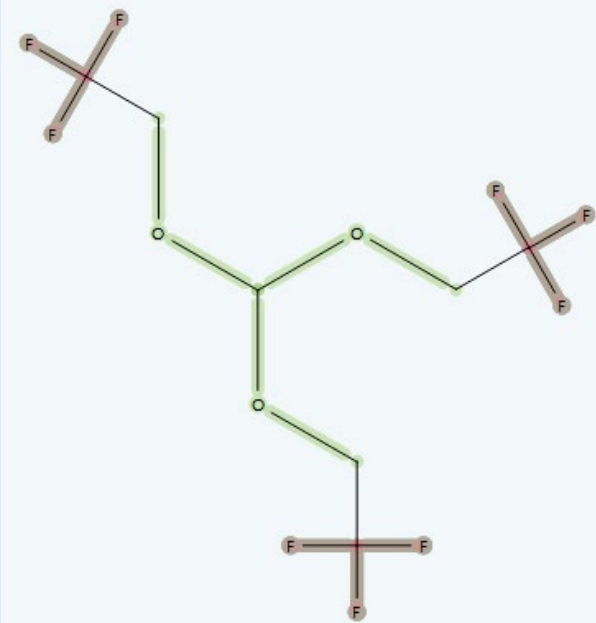
Molecular Fingerprints



MACCS



ECFP



Toxprints

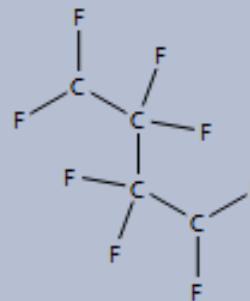
Output:

```
01000001011011100110010001110010011001010111011100100000010101110110100001100101011001010110110001100101
01110010001000000110100101110011001000000110000100100000011001100110000101110100001000000110001001101001
0111010001100011011010001
```

Toxprints

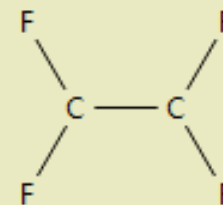
- 729 fragments
- PFAS substructures
- Good functional groups
- Some scaffolds

bond: CX_halide_alkyl-F_perfluoro_butyl 147



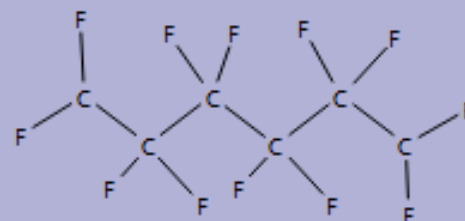
211

bond: CX_halide_alkyl-F_perfluoro_ethyl 148



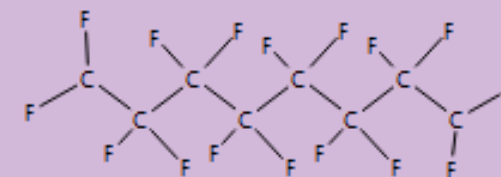
354

bond: CX_halide_alkyl-F_perfluoro_hexyl 149



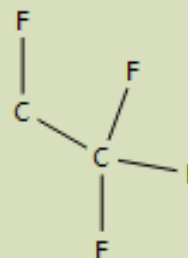
135

bond: CX_halide_alkyl-F_perfluoro_octyl 150



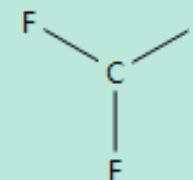
59

bond: CX_halide_alkyl-F_tetrafluoro_(1_1_1_2-) 151



283

bond: CX_halide_alkyl-F_trifluoro_(1_1_1-) 152

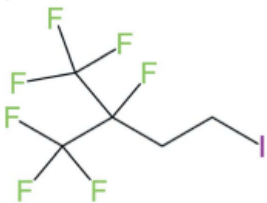
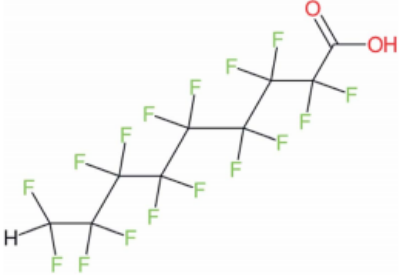
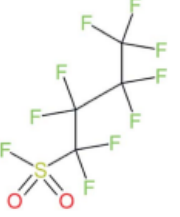




336

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.

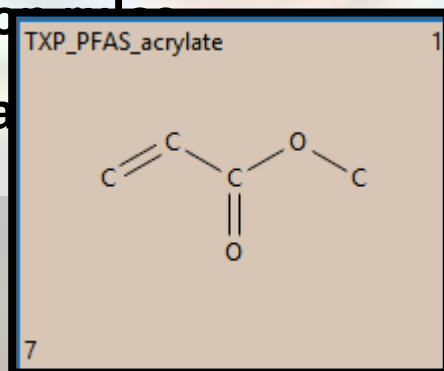
CAS_RN	Example structure	Explanation
Branched or cyclic perfluoroalkyl chains		
99324-96-6 Other examples: 28788-68-3 (ring)		This structure contains a branched perfluoroalkyl chain with two terminal CF ₃ groups. To capture these, the default "pacs" SMARTS may need adjusting in future studies. It is 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 _n F _{2n})-X-R, where here X = C(=O). Other members followed e.g. a Cl-(C _n F _{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 _n F _{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 _n F _{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 _n F _{2n+1})X-R-X(C _m F _{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

[illegible]

CSRML:

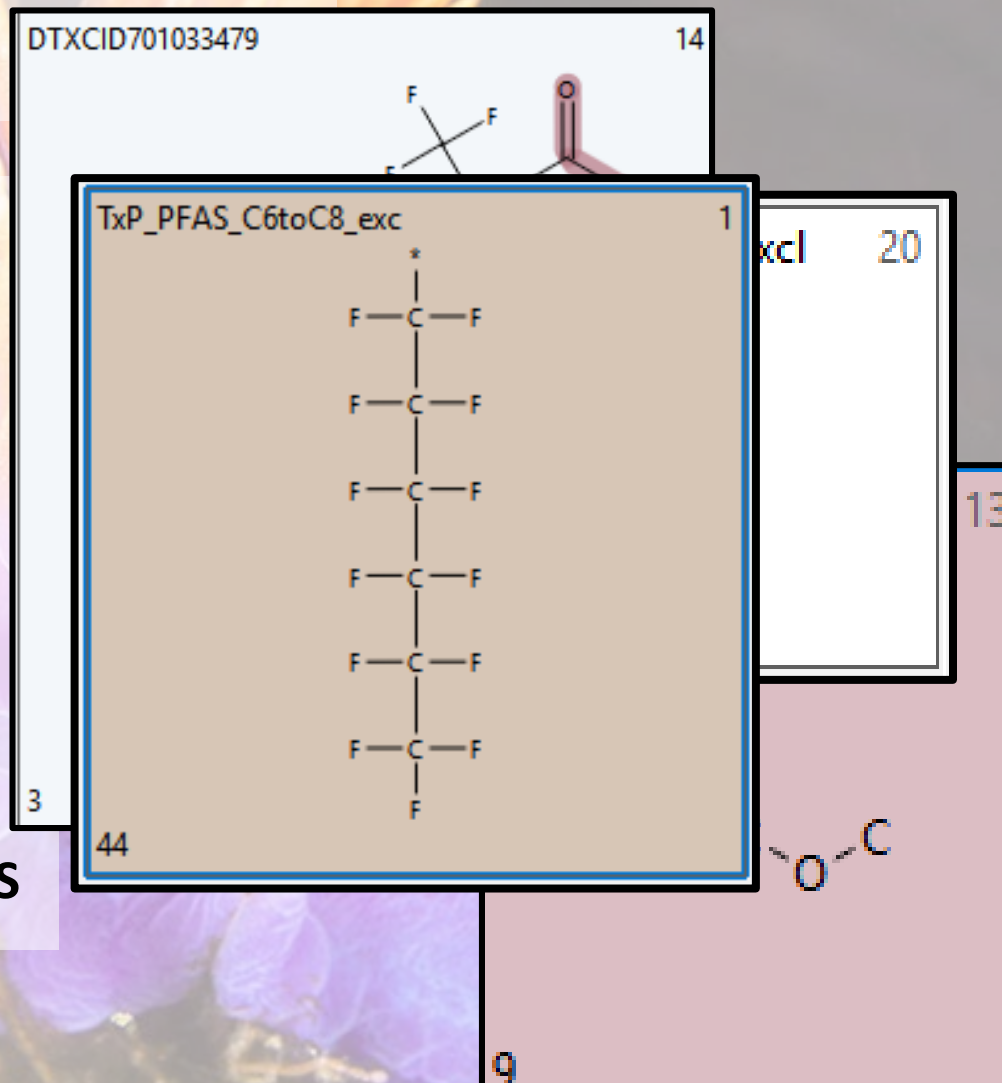
- Chemical Subgraphs and Reactions Markup Language (CSRML)
- XML based language
- Supports connectivity and topology but also properties of atoms, bonds, electronic systems
- Reaction networks
- Freeware



```
1005 > <classes id="ttc">=
1085 <subgraph id="TXP-000-0000-0000-0000-0000">
1086 NEW_TXP_PFAS.xml x toxprint_V2.0_r711.xml x PFAS_TXP_v1.xml x
1087 1 <?xml version='1.0' encoding='utf-8'?>
1088 2 <csrml xmlns="http://www.molecular-networks.com/schema/csrml" id="TxP_PFAS_Categories" csrmlVersion="2">
1089 3 <title>
1090 4 ToxPrint PFAS/PFOA Categories Version 1.0
1091 5 </title>
1092 6 <description>
1093 7 $Id: PFAS_Categories.xml Lougee $
1094 8 $Author: Lougee $
1095 9 </description>
1096 10
1097 11 > <classes id="TxP_PFAS_Categories">=
1098 49 <!-- Acrylate from scratch not really just m179-->
1099 50 <subgraph id="TxP_PFAS_acrylate">
1100 51 <label>TXP_PFAS_acrylate</label>
1101 52 <title>TXP_PFAS_acrylate</title>
1102 53 <comment>TXP_PFAS_acrylate</comment>
1103 54 <molecule id="m188">
1104 55 <matchIf feature="substructureMatch"/>
1105 56 <atoms>
1106 57 <atom element="C" x="1.7411" y="0.9999" id="a1"/>
1107 58 <atom element="O" x="2.6203" y="1.4998" id="a2"/>
1108 59 <atom element="O" x="1.7411" y="0.0" id="a3"/>
1109 60 <atom element="C" x="3.4823" y="0.9999" id="a4"/>
1110 61 <atom element="C" x="0.8792" y="1.4998" id="a5"/>
1111 62 <atom element="C" x="0.0" y="0.9999" id="a6"/>
1112 63 </atoms>
1113 64 <bonds>
1114 65 <bond order="single" id="b1">
1115 66 <atom id="a1"/>
1116 67 <atom id="a2"/>
1117 68 </bond>
1118 69 <bond order="double" id="b2">
1119 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**



PFAS_TXP_v1.xml +

Chemotype Sets

- ☒ TxP_PFAS_Categories
 - ☒ TxP_PFAS_COOR
 - ☒ TXP_PFAS_acrylate
 - ☒ TxP_PFAS_acylhalide
- ☒ TxP_PFAS_alcohol
 - ☒ TxP_PFAS_alcohol_polyF
 - ☒ TxP_PFAS_alcohol_primary
 - ☒ TxP_PFAS_alcohol_primary_FT_d...
 - ☒ TxP_PFAS_alcohol_primary_FTn1
 - ☒ TxP_PFAS_alcohol_primary_FTn2
 - ☒ TxP_PFAS_alcohol_sulfonylamide
- ☒ TxP_PFAS_aldehydeanhydride
- ☒ TxP_PFAS_alkylXprimary
- ☒ TxP_PFAS_alkylXtertiaryxCO
- ☒ TxP_PFAS_amine
 - ☒ TxP_PFAS_amine_ether
 - ☒ TxP_PFAS_amine_primary
- ☒ TxP_PFAS_carboxamide
- ☒ TxP_PFAS_ether
- ☒ TxP_PFAS_ethylene_xCO
- ☒ TXP_PFAS_ketone
- ☒ TxP_PFAS_oxidehydroxy
- ☒ TxP_PFAS_perFhexyl
- ☒ TxP_PFAS_perFoctyl
- ☒ TxP_PFAS_silane
- ☒ TxP_PFAS_sulfonyl
 - ☒ TxP_PFAS_sulfonamide
 - ☒ TxP_PFAS_sulfonamide_alcohol
 - ☒ TxP_PFAS_sulfonate
 - ☒ TxP_PFAS_sulfonate_FTn2
 - ☒ TxP_PFAS_sulfonylhalide

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

The screenshot displays the Chemotyper software interface. At the top, a menu bar includes File, Edit, View, Selection, Find, Packages, and Help. Below the menu, the main window shows a list of chemical structures with their IDs (DTXCID4035251, DTXCID701033479, DTXCID50113437) and corresponding chemical structures. A central error dialog box is open, displaying a red 'X' icon and the message: "The specified XML file is not a valid CSRML file: C:\Users\Administrator\OneDrive\Profile\Desktop\TXP_PFAS_v1.6.xml". The error details section shows: "Error at file 'C:\Users\Administrator\OneDrive\Profile\Desktop\TXP_PFAS_v1.6.xml', line 31976, column 41. Message: no character data is allowed by content model". To the right of the error dialog, a sidebar shows a list of chemical features with checkboxes, including TxP_PFAS_Q6_ring, TxP_PFAS_Q7_ring, TxP_PFAS_Q8_ring, TxP_PFAS_Q9_ring, Bicyclo Rings, Chain Triple, Misc Functional Groups, TxP_PFAS_alkyne, TxP_PFAS_amine, TxP_PFAS_imino, TxP_PFAS_nitrile, TxP_PFAS_nitro, TxP_PFAS_phosphate, and TxP_PFAS_sulfonyl. The bottom of the screen shows a code editor with XML code:

```
31987 <atom element="QRY" id="a5">
31988 <matchIf feature="atomList">
```

Some Things I Like

ChemoTyper

Menu

Welcome

Browse

Match

PFAS_test2.sdf

DTXCID8027583 7

DTXCID2039041 9

DTXCID4035251 12

DTXCID20330876 19

DTXCID20897070 20

DTXCID9027578 22

1 / 26

Filter Structures by ID type ID Filter Pattern

Filter Chemotypes No Filter

Structures Loaded: 147 Total Coverage: 147 Selected: 0 Matched: 26 ID: NAME

TXP_PFAS_v1.6.xml

Chemotype Sets

- Perfluoro Chain Length Exclusive
 - ☐ TxP_PFAS_C1_excl
 - ☐ TxP_PFAS_C2_excl
 - ☐ TxP_PFAS_C3_excl
 - ☒ TxP_PFAS_C4_excl
 - ☐ TxP_PFAS_C5_excl
 - ☐ TxP_PFAS_C6_excl
 - ☐ TxP_PFAS_C7_excl
 - ☐ TxP_PFAS_C8_excl
 - ☐ TxP_PFAS_C9_excl
 - ☐ TxP_PFAS_C10_excl
 - ☐ TxP_PFAS_C11_excl
 - ☐ TxP_PFAS_C12_excl
 - ☐ TxP_PFAS_C13_excl
 - ☐ TxP_PFAS_C14_excl
 - ☐ TxP_PFAS_C15_plus
- Perfluoro Chain Length Exclusive Un...
 - ☐ TxP_PFAS_C1_nocap_excl
 - ☐ TxP_PFAS_C2_nocap_excl
 - ☐ TxP_PFAS_C3_nocap_excl
 - ☒ TxP_PFAS_C4_nocap_excl
 - ☐ TxP_PFAS_C5_nocap_excl
 - ☐ TxP_PFAS_C6_nocap_excl
 - ☐ TxP_PFAS_C7_nocap_excl
 - ☐ TxP_PFAS_C8_nocap_excl
 - ☐ TxP_PFAS_C9_nocap_excl
 - ☐ TxP_PFAS_C10_nocap_excl
 - ☐ TxP_PFAS_C11_nocap_excl
 - ☐ TxP_PFAS_C12_nocap_excl
 - ☐ TxP_PFAS_C13_nocap_excl

2 / 2

Filter Chemotypes by ID type ID Filter Pattern

Filter Structures Containing Any Selected Chemotype (OR)

Chemotypes Loaded: 143 Total Coverage: 86 Selected: 2 (110 hidden) ID: Auto

TxP_PFAS_C4_excl

17

TxP_PFAS_C4_nocap_excl

9

How to use these now?

ChemoTyper

PFAS_test2.sdf +

DTXCID9039369 48 DTXCID3040061 49 DTXCID80577861 50 DTXCID6034390 51

DTXCID5035589 52 DTXCID801021863 53 DTXCID601030672 54 DTXCID90126980 55

DTXCID80331283 56 DTXCID60331822 57 DTXCID9038713 58 DTXCID1039056 59

48 / 147

Filter Structures by ID type ID Filter Pattern

Filter Chemotypes No Filter

Structures Loaded: 147 Total Coverage: 147 Selected: 0 ID: NAME

TXP_PFAS_v1.6.xml +

ChemoType Sets

- ☒ PFAS Toxprint Categories
 - ☒ TxP_PFAS_generic_CF2_CF
 - ☒ TxP_PFAS_generic_CF_chain
 - ☒ TxP_PFAS_generic_C2F4
 - ☒ Perfluoro Chain Length Exclusive
 - ☒ Perfluoro Chain Length Exclusive U...
 - ☒ TxP_PFAS_generic_CF_ring
 - ☒ Bicyclo Rings
 - ☒ Carbon Rings
 - ☒ Fluorinated Carbon Rings
 - ☒ General Rings
 - ☒ TxP_PFAS_polyF_generic
 - ☒ Branching
 - ☒ Chain Double
 - ☒ Chain Quads and Above
 - ☒ Chain Triple
 - ☒ Functionalization Categories
 - ☒ Carbon Bonds
 - ☒ Fluorotelomer-type
 - ☒ TxP_PFAS_alkene
 - ☒ TxP_PFAS_alkene_ether
 - ☒ TXP_PFAS_alkyne
 - ☒ Nitrogen-Based Functionalization
 - ☒ Oxygen-Based Functionalization
 - ☒ Phosphate-Based Functionalization
 - ☒ Silicon-Based Functionalization
 - ☒ Sulfur-Based Functionalization
 - ☒ TxP_PFAS_inorganic_F
 - ☒ TxP_PFAS_other_halogens

TxP_PFAS_generic_CF2_2 CF 136

TxP_PFAS_generic_CF_c hain 144

TxP_PFAS_generic_CF_ring 7

TxP_PFAS_C1_excl 32

TxP_PFAS_C2_excl 9

TxP_PFAS_C3_excl 20

TxP_PFAS_C4_excl 17

TxP_PFAS_C5_excl 4

2 / 143

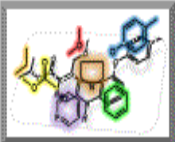
Filter Chemotypes by ID type ID Filter Pattern

Filter Structures No Filter

Chemotypes Loaded: 143 Total Coverage: 86 Selected: 143 ID: Auto

How to use these now?

[Register](#) | [Login](#)



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

Advanced Search

Backlogs

[Bug Backlog](#)

[Feature Requests](#)

Wiki

[Wiki Homepage](#)

ToxPrint Chemotypes

[Acknowledgement](#)

The ChemoType 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)

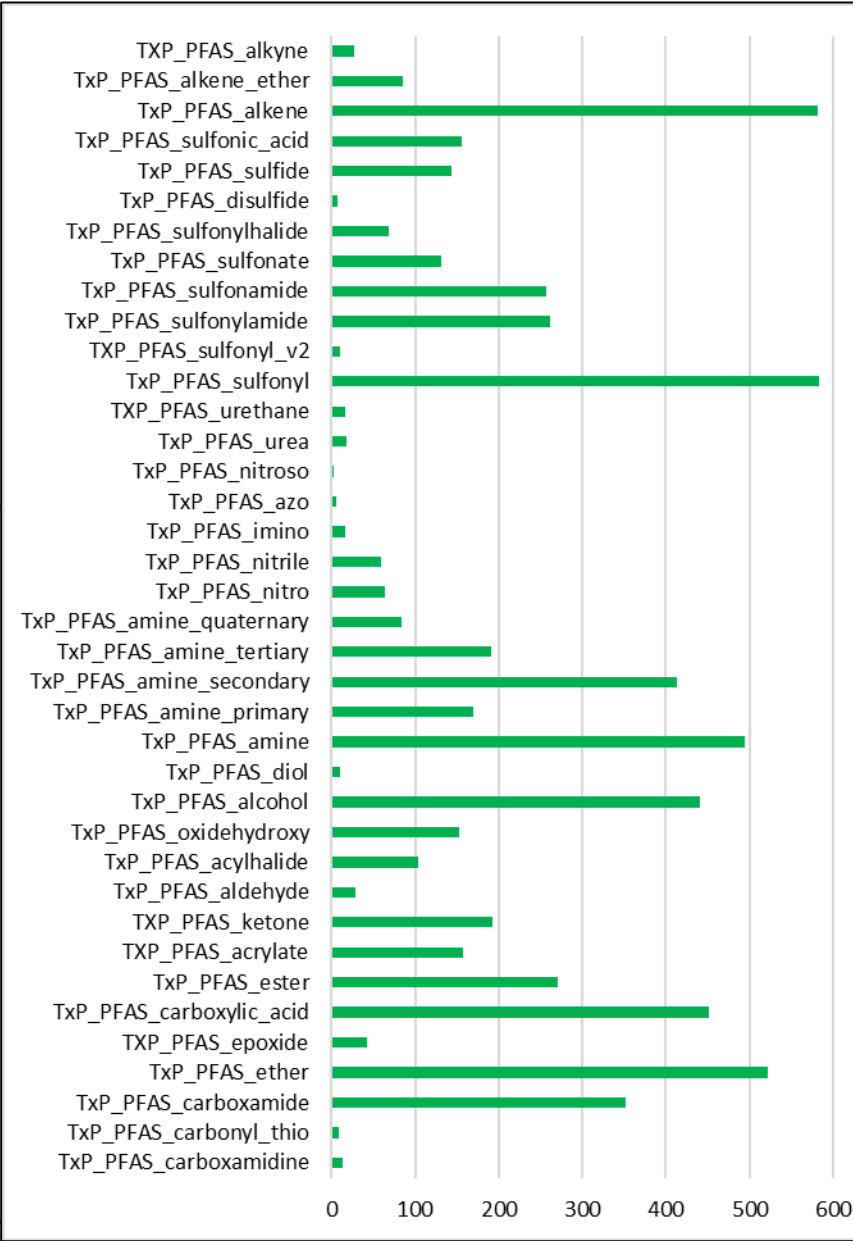
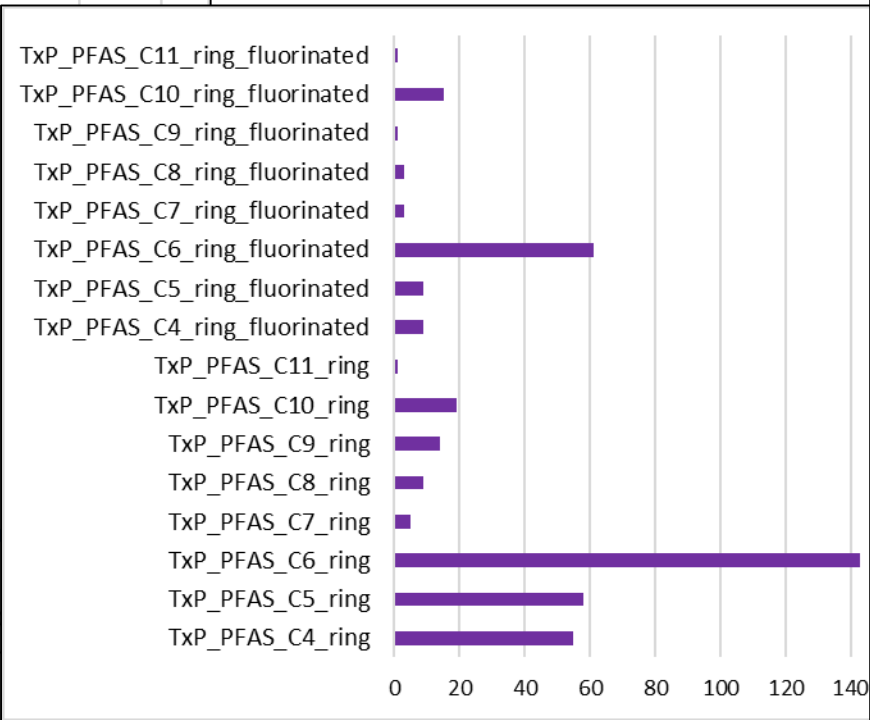
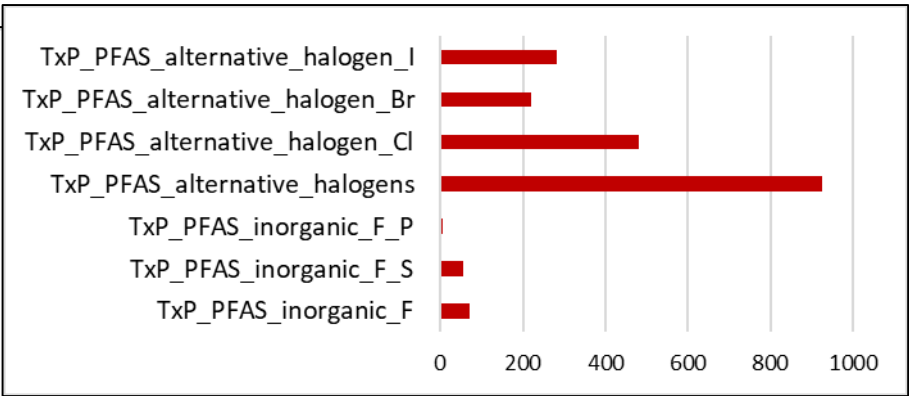
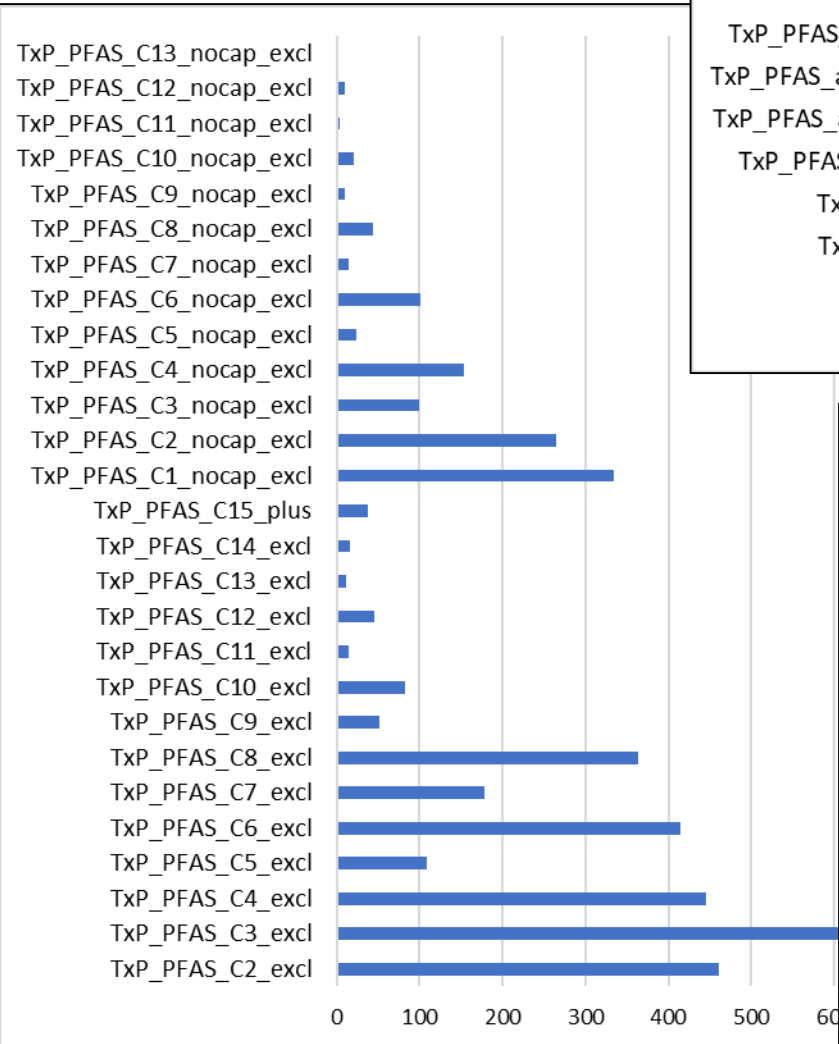
Structural Rules and Alerts

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

