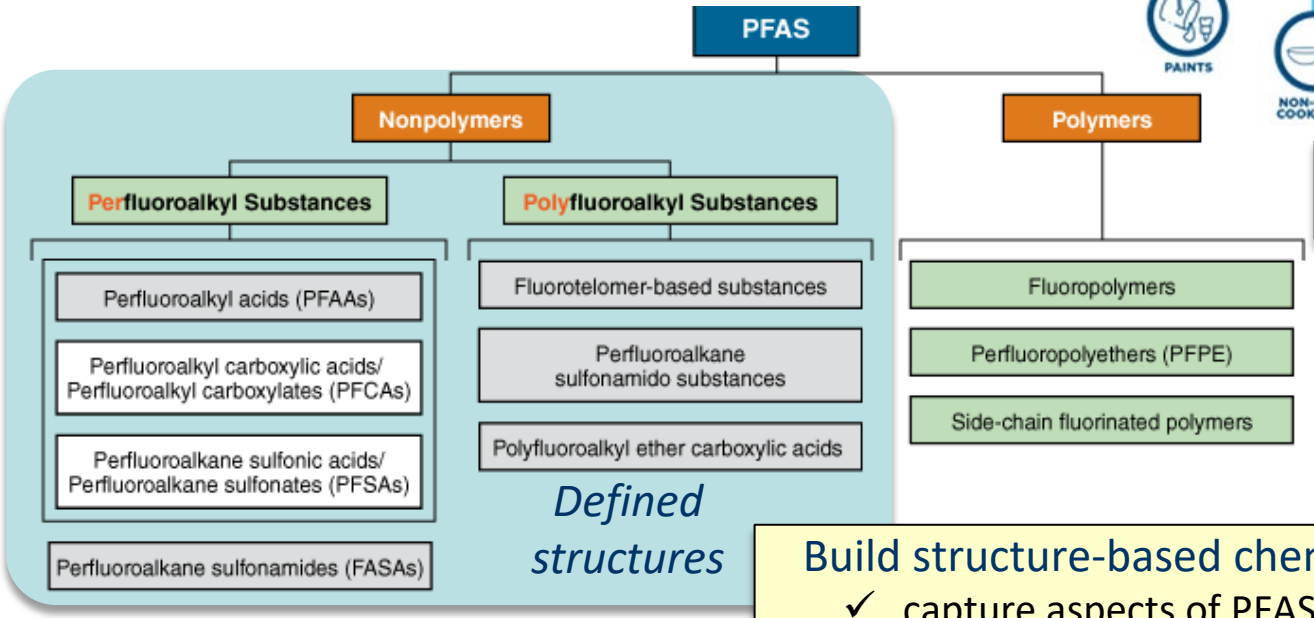
New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Abstract #263

Ann Richard[1], Ryan Lougee[2], Grace Patlewicz[1], Chris Grulke[1], Antony Williams[1], Chihae Yang[3], James Rathman[4], Tomasz Magdziarz[3]
[1]Ctr. for Comput. Toxicol. & Exposure, ORD, U.S. EPA; [2] ORISE, Oak , TN; [3]MN-AM, Nuremberg, Germany, [4]MN-AM, Columbus, OH

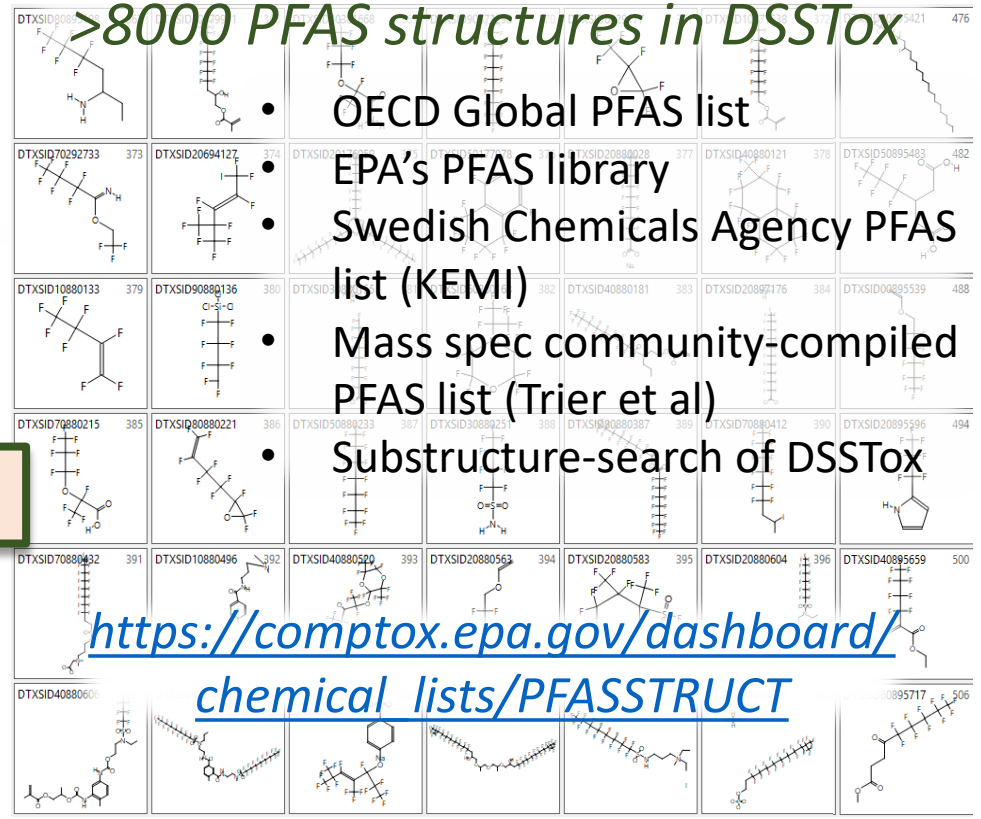
- Per- and polyfluoroalkyl substances (PFAS) are of concern due to widespread production, environmental persistence, and adverse ecological and health impacts (*Lau et al., 2007*)
- Expert-based PFAS categories and standardized naming terminologies are difficult for non-experts to apply and ill-suited to describing the broad diversity of PFAS structures



Problem

>8000 PFAS structures in DSSTox

- OECD Global PFAS list
- EPA's PFAS library
- Swedish Chemicals Agency PFAS list (KEMI)
- Mass spec community-compiled PFAS list (Trier et al)
- Substructure-search of DSSTox



https://comptox.epa.gov/dashboard/chemical_lists/PFASSTRUCT

Objectives

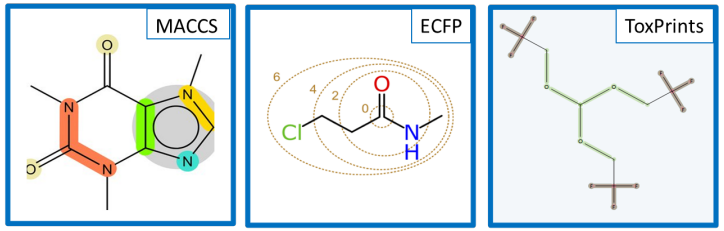
Build structure-based chemical features to fingerprint and categorize PFAS that:

- ✓ capture aspects of PFAS chemistry potentially impacting reactivity, bioactivity, fate & transport
- ✓ are chemically intuitive and easy to use to profile/categorize new and existing PFAS chemicals
- ✓ are reproducible and amenable to automation and cheminformatics application
- ✓ are publicly available, visualizable, and accessible to chemists and non-chemists

New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Approach: Building & Validating the PFAS ToxPrint Feature Set

Publicly available molecular fingerprinting methods:



Binary Output:
Chem1: 01000001011011100110010001110010011001010111011100100000
Chem2: 01000001100001001000000110011001100001011101000010000001

CSRML: Chemical Subgraph & Reaction Mark-up Language

- XML-based, can create hierarchical listing of features in public Chemotyper
- Can represent generic query features without enumeration (unlike SMARTS)
- Can specify chain lengths, range of chain lengths, & multiple fragment conditions in a single feature
- Can be extended to include atom and bond properties

Advantages of ToxPrints (coded in open CSRML)

- 729 structural features spanning diverse chemical space
- Chemically informative and intuitive names
- Good coverage of functional groups, includes some PFAS substructures
- Can visualize and export from publicly available Chemotyper

→ ***BUT,*** *ToxPrints are missing many important PFAS concepts*

Create CSRML-based TxP_PFAS set

- ✓ PFAS Terminology paper - expert categories (*Buck et al., 2011; Cousins et al., 2020*)
- ✓ Collected structures related to toxicity & adverse outcomes
- ✓ Searched literature for interesting byproducts and structures
- ✓ Incorporated OECD PFAS Global list categories & structures (*Wang, 2018*) as well as missing categories (*Sha et al., 2019*)
- ✓ Incorporated/modified functional groups from public ToxPrint CSRML file (*Yang et al., 2015*)
- ✓ Created new PFAS CSRML features, validated in Chemotyper with structures from PFASSTRUCT and PFASOECD lists (https://comptox.epa.gov/dashboard/chemical_lists)
- ✓ Added PFAS-defining features (i.e., bounded PFAS space of PFASSTRUCT) and generic features to capture broader category concepts

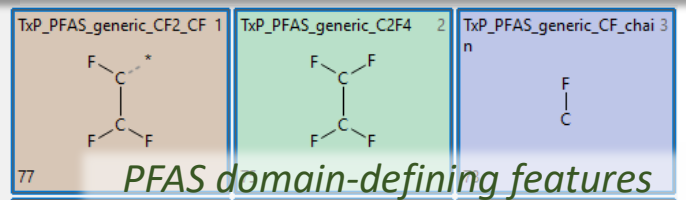
New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Results: TxP_PFAS CSRML file

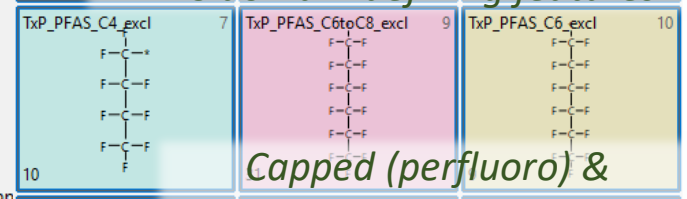
TXP_PFAS_v1.11.4.xml

Chemotype Sets

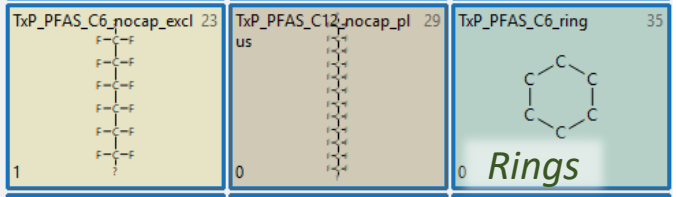
- PFAS Toxprint Categories
  - Structural Categories
    - TxP_PFAS_generic_CF2_CF
    - Chains
      - TxP_PFAS_generic_C2F4
      - TxP_PFAS_generic_CF_chain
      - Perfluoro Chain Length Exclusive
        - TxP_PFAS_C1_excl
        - TxP_PFAS_C2_excl
        - TxP_PFAS_C3_excl
        - TxP_PFAS_C4_excl
        - TxP_PFAS_C5_excl
        - TxP_PFAS_C6toC8_excl
        - TxP_PFAS_C9_plus
        - TxP_PFAS_C12_plus
      - Perfluoro Chain Length Exclusive Uncapp
    - Rings
      - TxP_PFAS_aromatic_ring_generic_C2F2
      - TxP_PFAS_generic_CF_ring
      - Carbon Rings
        - TxP_PFAS_C4_plus_rings
      - General Rings
        - TxP_PFAS_Q4_plus_rings
      - Bicyclo Rings
        - TxP_PFAS_bicyclo_Q6--Q5_ring
        - TxP_PFAS_bicyclo_Q6--Q6_ring
        - TxP_PFAS_bicyclo_Q6-Q5_ring
        - TxP_PFAS_bicyclo_Q6-Q6_ring
        - TxP_PFAS_bicyclo_Q6Q5_ring
        - TxP_PFAS_bicyclo_Q6Q6_ring
    - Polyfluoro
      - TxP_PFAS_polyF_generic
      - TxP_PFAS_polyF_cap_CF2H
      - TxP_PFAS_polyF_cap_CFH2
      - TxP_PFAS_polyF_chain_CCHFC
      - TxP_PFAS_polyF_chain_CF2CHFCF2
      - TxP_PFAS_polyF_chain_CFCH2CF
    - Branching
      - Chain Double
        - TxP_PFAS_C1_excl_double
        - TxP_PFAS_C1_plus_double
        - TxP_PFAS_C2_plus_double
        - TxP_PFAS_perF_isopropyl
      - Chain Quads and Above
      - Chain Triple



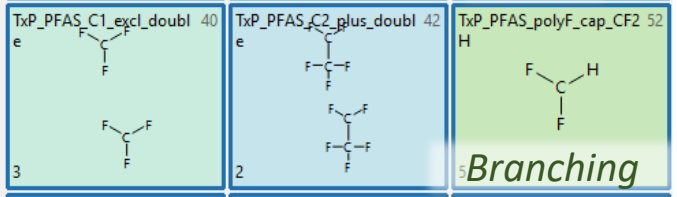
PFAS domain-defining features



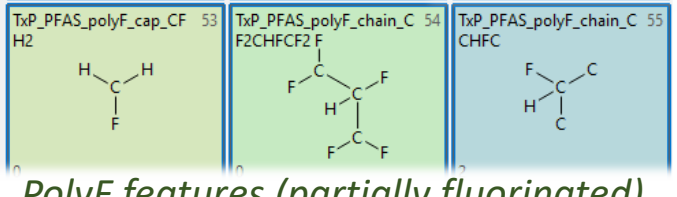
Capped (perfluoro) & uncapped perfluoro chains



Rings



Branching

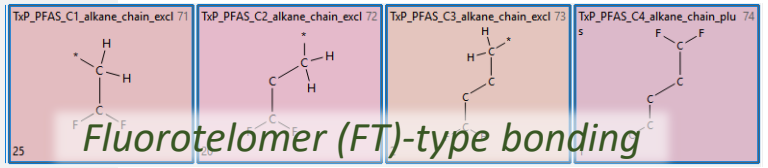


PolyF features (partially fluorinated)

Structural Categories

Functional Categories

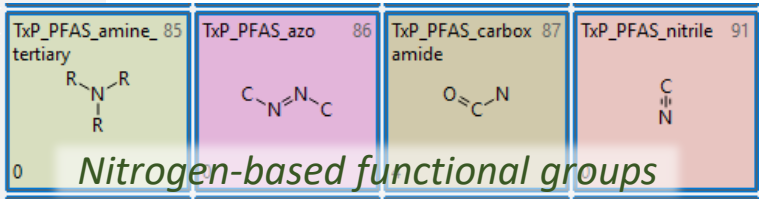
- Functional Categories
  - Carbon Bonds
    - Fluorotelomer-type
    - TxP_PFAS_alkene
    - TxP_PFAS_alkene_ether
    - TxP_PFAS_alkyne
  - Metals
  - Nitrogen-Based
    - TxP_PFAS_nitro
    - TxP_PFAS_nitroso
    - TxP_PFAS_urea
    - TxP_PFAS_urethane
  - Oxygen-Based
    - Carbonyl
      - TxP_PFAS_acylhalide
      - TxP_PFAS_aldehyde
      - TxP_PFAS_anhydride
      - TxP_PFAS_carboxylic_acid
      - TxP_PFAS_ketone
    - TxP_PFAS_acrylate
    - TxP_PFAS_alcohol_aliphatic
    - TxP_PFAS_ester
    - TxP_PFAS_ether
  - Phosphate-Based
    - TxP_PFAS_organophosphine
    - TxP_PFAS_phosphate
    - TxP_PFAS_phosphinic_acid
    - TxP_PFAS_phosphonate
  - Silicon-Based
  - Sulfur-Based
    - TxP_PFAS_sulfide
      - TxP_PFAS_carbonyl_thio
      - TxP_PFAS_disulfide
      - TxP_PFAS_sulfonyl
      - TxP_PFAS_thioester
      - TxP_PFAS_thioketone
    - TxP_PFAS_inorganic_F
    - TxP_PFAS_other_halogens
      - TxP_PFAS_alkylXtertiary
      - TxP_PFAS_other_halogen_Br



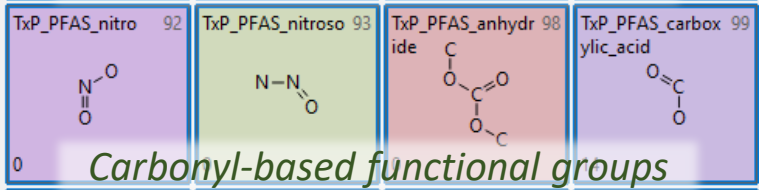
Fluorotelomer (FT)-type bonding

138 Total TxP_PFAS Features:

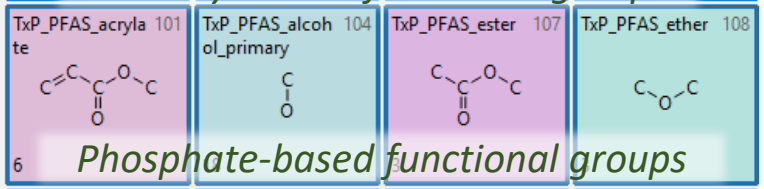
- *All PFASSTRUCT & PFASOECD structures contain ≥ 1 TxP_PFAS*
- *All TxP_PFAS are represented in both inventories*



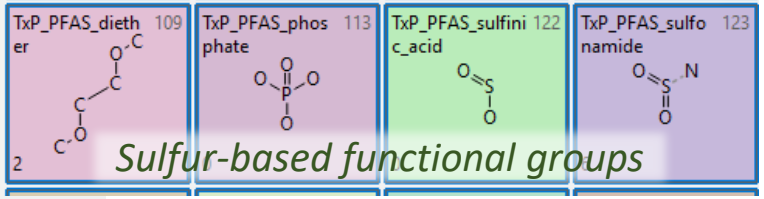
Nitrogen-based functional groups



Carbonyl-based functional groups



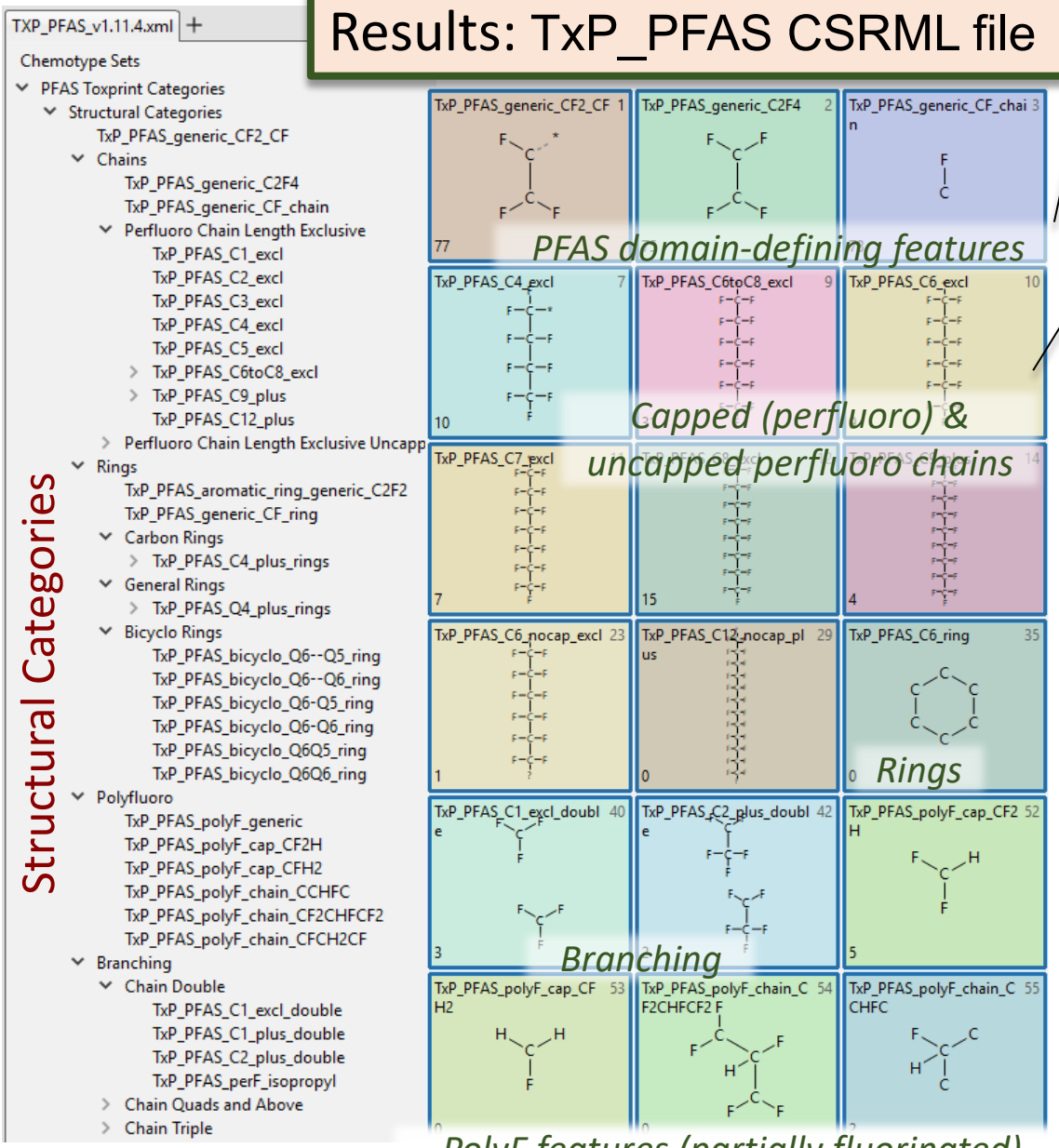
Phosphate-based functional groups



Sulfur-based functional groups

New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Results: TxP_PFAS CSRML file



Features defining the PFASSTRUCT space:
must be present for chemical to be defined as PFAS

"excl" ….exclusive chain length (e.g., only C6 chain length)
"plus" ….includes all higher chain lengths (e.g., C9 and above)
"cap" ….terminal group
"nocap"...open ended terminal group
"polyF" …incomplete fluorination (i.e., some C-H bonds)

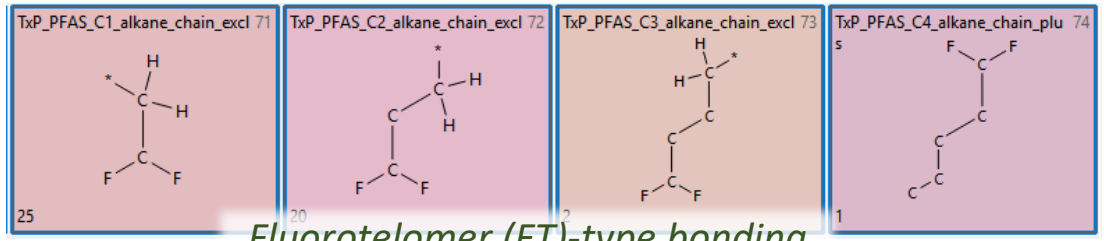
PFASSTRUCT (8150) Sample "Category Totals"

- 1886 perfluoro ≥ C6 substances
  - ➢ 974 ≥ C8 perfluoro
    - ➢ 474 fluorotelomer-type
      - ➢ 95 sulfonyl-containing (acid, ester, sulfonamide)
- 421 uncapped perfluoro ≥ C6 substances
- 338 perfluoro rings ≥ C6
- 1862 branched perfluoro (C1-C2)
  - ➢ 269 branched + perfluoro ≥ C6 substances

TxP_PFAS features alone or in combinations provide flexible means for constructing structure-based PFAS category definitions.

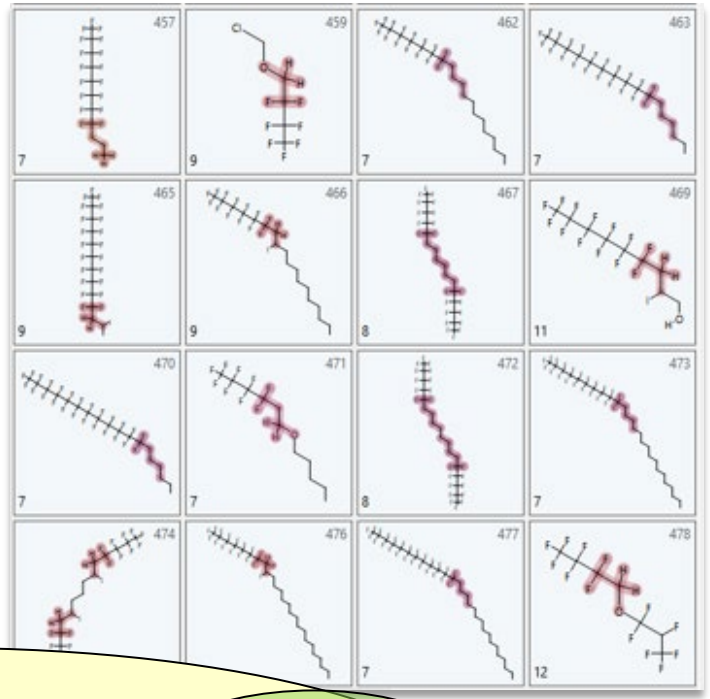
New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Application: TxP_PFAS vs. Name-based categorization



Fluorotelomer (FT)-type bonding

- Fluorotelomers (FTs) that contain PFOA precursors can be metabolized into, and degrade to, PFOA
- PFOA is also formed as byproduct in production of FTs, and is found in goods treated with fluorotelomers, including food contact substances.



Do fluorotelomer-type TxP_PFAS features identify expert-assigned fluorotelomer-named compounds?

PFASSTRUC	8150 structures	122 FT names
TxP_PFAS_C1_alkane_chain_excl	951	10
TxP_PFAS_C2_alkane_chain_excl	725	96
TxP_PFAS_C3_alkane_chain_excl	90	5
TxP_PFAS_C4_alkane_chain_plus	191	4
Structure contains Fluorotelomer-type TxP	1957	115

1957 /8150 **PFASSTRUCT** contain FT-type TxP (115 /122 match FT name)

122 contain "fluorotelomer" in name

- *Name-based FT category is PFAS domain-specific, requires chemical expertise, and FT names are not consistently applied nor universally understood*
- *TxP_PFAS FT-type features support a clear, reproducible, easy to apply, structure-based category representation*

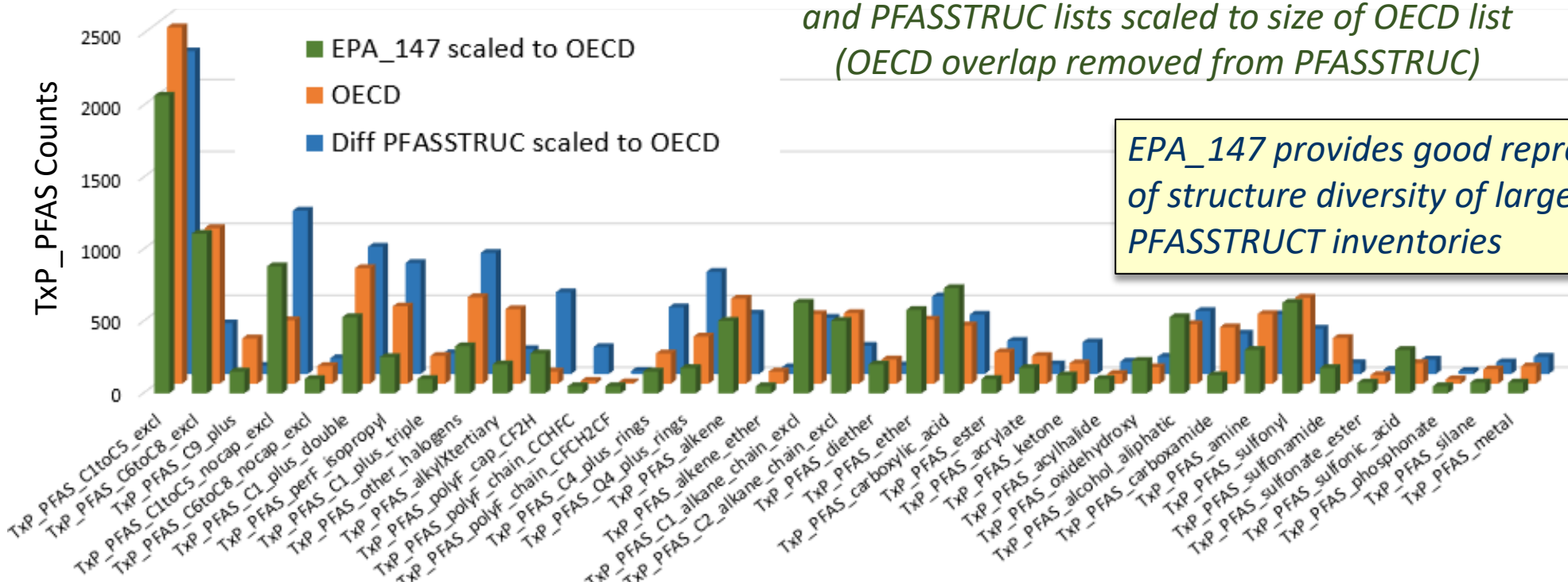
New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Application: PFAS Inventory Profiling & EPA_147 PFAS screening

EPA_147 is set of 147 PFAS chemicals undergoing Tier 1 screening within EPA's ToxCast program

- Subset of 34 TxP_PFAS were used to categorize EPA's inventory for chemical selection, read-across and assay data analysis (*Patlewicz et al., 2019*)
- 88 total TxP_PFAS represented in 147 set

Compare EPA_147 testing inventory to OECD and PFASSTRUC lists scaled to size of OECD list (OECD overlap removed from PFASSTRUC)



EPA_147 provides good representation of structure diversity of larger OECD and PFASSTRUCT inventories

New CSRML-based features to categorize and fingerprint PFAS structure lists for cheminformatics analysis and read-across

Summary

TxP_PFAS CSRML:

- Captures chemically relevant features characterizing known, structurally diverse PFAS inventory
- Enables consistent, reproducible, structure-based profiling and grouping of PFAS chemicals
- Can be imported into public Chemotyper (https://chemotyper.org) to visualize and group PFAS structures with features
- Will enable standardized exchange of PFAS information based on structure

Future Plans

- Make TxP_PFAS CSRML file publicly available on the ToxPrint website, https://toxprint.org/ and TxP_PFAS fingerprint file for PFASSTRUCT on https://figshare.com/
- Apply TxP_PFAS to modeling and read-across of assay data to identify features associated with bioactivity
- Establish TxP_PFAS feature correspondence with widely used PFAS category concepts (alone or in combination with other features)
- Add new TxP_PFAS features/properties as needed to capture evolving understanding of PFAS category patterns in relation to bioactivity and environmental fate & transport

References

1. Lau, C., et al., Perfluoroalkyl acids: a review of monitoring and toxicological findings. Toxicol Sci, 2007. 99(2): p. 366-94.
2. Buck, R.C., et al., Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. Integr Environ Assess Manag, 2011. 7(4): p. 513-41.
3. Cousins, I.T., et al., Strategies for grouping per-and polyfluoroalkyl substances (PFAS) to protect human and environmental health. Environ Sci: Processes & Impacts, 2020.
4. Wang, Z., S25| OECDPFAS| List of PFAS from the OECD. Zenodo Dataset, 2018, DOI 10.5281/zenado.2648776.
5. Sha, B., et al., Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs). Environ Sci Process Impacts, 2019. 21(11): p. 1835-1851.
6. Yang, C., et al., New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. J Chem Inf Model, 2015. 55(3): p. 510-28.
7. Patlewicz, G., et al., A chemical category-based prioritization approach for selecting 75 per-and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. Environ Health Perspec, 2019, 127(01), 014501.