New public CSRML-based structure-fingerprint method for profiling and categorizing PFAS structures for modeling and read-across

March 12-26, 2021
SOT Meeting Virtual

Ann Richard[1], Ryan Lougee[2], Grace Patlewicz[1], Christopher Grulke[1], Antony Williams[1], Chihae Yang[3], James Rathman[4], Tomasz Magdziarz[3]

[1] Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[2] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN 37831, USA
[3] MN-AM, Nürnberg, Germany
[4] MN-AM, Columbus, OH 43215 USA

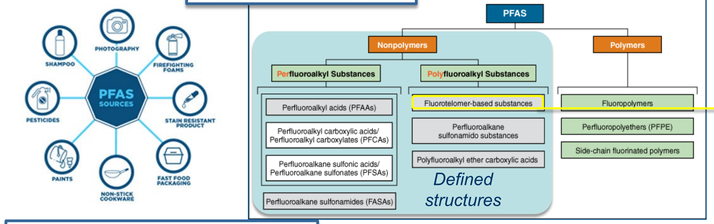
Contact: richard.ann@epa.gov

www.epa.gov

Abstract

Per- and polyfluoroalkyl substances (PFAS) are of high public interest due to widespread production, environmental persistence, and adverse ecological and health impacts. EPA's CompTox Chemicals Dashboard has published over 8000 curated PFAS structures, which encompasses the structurable content of several public PFAS lists, including the OECD PFAS list *(Wang et al., 2018)*. Whereas most studies to-date have focused on the health effects of a small number of PFAS compounds, such as PFOA and PFOS *(Lau et al., 2007)*, relatively little is known about the health effects of the vast majority of PFAS and their byproducts. Methods for profiling the PFAS chemical structure space are needed to support modeling and structure-based categorization efforts. However, naming conventions and publicly available molecular fingerprinting methods are ill-suited to capturing the wide range of potentially relevant PFAS structural patterns. Expert-defined PFAS chemical category terms are limited to simpler, single functional categories (e.g., perfluorocarboxylic acids) and often lack clear structure definition *(Buck et al., 2011)*. Using the publicly available CSRML (Chemical Subgraphs and Reactions Markup Language) *(Yang et al., 2016)*, we developed a set of 138 PFAS ToxPrint features, which includes an expanded set of ToxPrint functional groups (https://toxprint.org/), augmented by 74 new PFAS fingerprints capturing category concepts, as well as important aspects of PFAS structures, including perfluoro chains, polyfluoro substructures, fluorinated rings, and various perfluoro branching patterns. These CSRML PFAS categories and features can be processed with the public Chemotyper (https://chemotyper.org/), provide comprehensive coverage of available PFAS lists, and are being used to profile and categorize PFAS chemical lists currently undergoing testing within EPA.

Types of PFAS Categories



Why do we need PFAS categories?

- Use categories share chemical properties impacting exposure and bioactivity (e.g., surfactants) (*Cousins et al., 2020*)
- Structure categories may exhibit similar properties & toxicity, e.g., C6-C8 perfluoro acids
- PFAS structural elements (such as incomplete fluorination, branching) can affect reactivity, ADME properties, etc.
- Presence of certain functional groups, e.g., acid groups, sulfonyls, and phosphates, may confer similar properties
- Commonly used for regulatory groupings and provide support for "Read-across" approaches

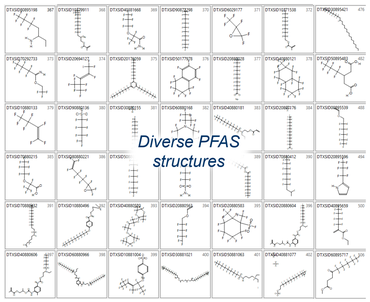
Past Approaches & Challenges

Standardized, terminology-based chemical-category naming approaches

- Expertise required, difficult to enforce standards
- Difficult to categorize diverse PFAS chemicals with multi-functional groups, branching, incomplete fluorination, etc. *(Buck et al., 2011)*

Structure-based approaches

- Current structure fingerprinting methods do not capture important structural concepts within PFAS space
- Capturing & standardizing general concepts related to branching, partial fluorination, etc. with current cheminformatics tools is difficult (*Sha et al., 2019*)

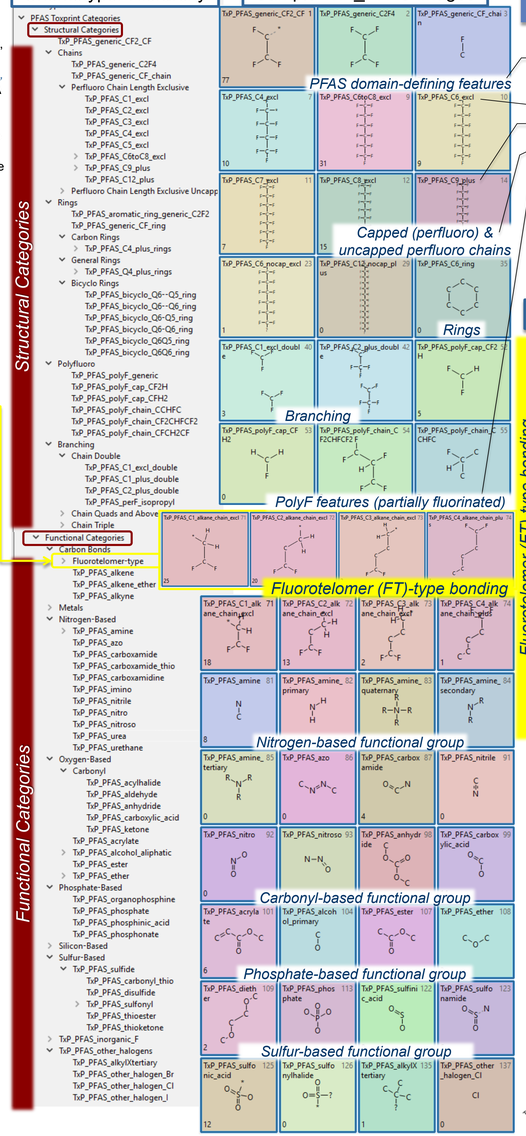


Objectives

Build structure-based chemical features for use in fingerprinting, profiling & categorizing PFAS that:

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

Chemotyper Hierarchy & Sample TxP_PFAS images



Results

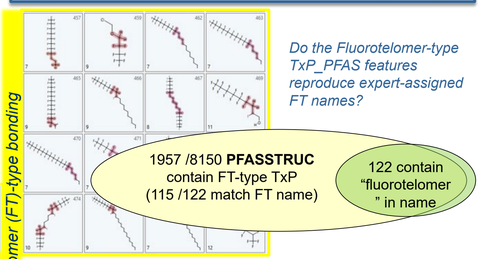
Bounding features of PFASSTRUC space, i.e., one or more must be present for chemical to be defined as PFAS

"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)

138 Total TxP_PFAS Features:

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

TP_PFAS vs. Name-based categorization: *e.g., FT-type*



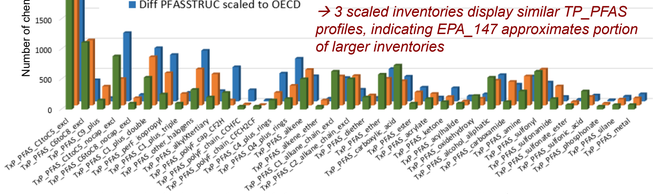
Do the Fluorotelomer-type TxP_PFAS features reproduce expert-assigned FT names?

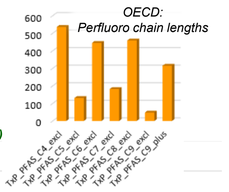
PFASSTRUC	8150 structures	122 FT names
TxP_PFAS_C1_alkane_chain_excl	951	10
TxP_PFAS_C2_alkane_chain_excl	725	96
TxP_PFAS_C3_alkane_chain_excl	90	5
TxP_PFAS_C4_alkane_chain_plus	191	4
Structure contains Fluorotelomer-type TxP	1957	115

→ *Name-based FT category is expert-based, provides incomplete coverage*
→ *TxP_PFAS FT-type features provide generalized structure-based, reproducible category representation*

PFAS Inventory Profiling

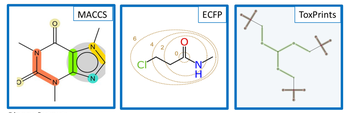
Compare counts of TxP_PFAS features contained in EPA_147 across 3 inventories (removing OECD overlap from PFASSTRUC, and scaled to size of OECD)





Approach: Building& Validating the PFAS ToxPrint Feature Set

Publicly available molecular fingerprinting methods:



Binary Output:
Chem1: 01000000101101110011001000111001001100101011011101100000000
Chem2: 01000000110000100100000011001100011000101110100001000001

Advantages of ToxPrints (coded in open CSRML)

- 729 structural features spanning diverse chemical space
- Names are chemically informative and intuitive
- Good coverage of functional groups, includes some PFAS substructures
- Can visualize and export from publicly available Chemotype

→ *BUT, do not adequately capture many important PFAS concepts*

CSRML (Chemical Subgraphs and Reactions Markup Language)

- XML based language, provides unique representations of features (unlike SMARTS)
- Can provide hierarchical organization of features in public Chemotyper
- Can specify chain lengths, range of chain lengths, and can include multiple fragment conditions in a single feature

Create CSRML-based PFAS_ToxPrint set

- ✓ PFAS Terminology paper - expert categories (*Buck et al., 2011*)
- ✓ Collected structures related to toxicity & adverse outcomes
- ✓ Searched literature for interesting byproducts and structures
- ✓ OECD PFAS Global list categories & structures (*Wang et al., 2018*)
- ✓ Noted missing OECD categories in *Sha et al., 2019*
- ✓ Incorporated and/or modified some features from ToxPrint CSRML file (e.g., functional groups)
- ✓ Created new PFAS features in CSRML, validated in Chemotyper with structures from PFASSTRUC and PFASOECD list (https://comptox.epa.gov/dashboard/chemical_lists)
- ✓ Added PFAS-defining features (bounding the PFAS space of PFASSTRUC) and generic features to capture broader category concepts

Conclusions and Future Plans

- Make CSRML file publicly available on the ToxPrint website, https://toxprint.org/
- Make full PFAS TxP fingerprint file available for PFASSTRUC on https://figshare.com/
- Apply TxP_PFAS to modeling and read-across of PFAS data using subsets or combinations of TxP_PFAS features
- Establish TxP_PFAS feature correspondence with widely used PFAS category concepts (alone or in combination with other features)
- Add new TxP_PFAS features as needed

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. Wang, Z., S25| OECDPFAS| List of PFAS from the OECD. Zenodo Dataset, DOI, 2018. 10.
4. Cousins, I.T., et al., Strategies for grouping per-and polyfluoroalkyl substances (PFAS) to protect human and environmental health. Environmental Science: Processes & Impacts, 2020.
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.

U.S. Environmental Protection Agency
Office of Research and Development