# Per- and polyfluoroalkyl substances (PFAS) in PubChem: 7 million and growing

Emma L. Schymanski[1§]*, Jian Zhang[2], Paul A. Thiessen[2], Parviel Chirsir[1], Todor Kondic[1], Evan E. Bolton[2§]*

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367 Belvaux, Luxembourg. ORCIDs: ELS: 0000-0001-6868-8145, PC: 0000-0002-9932-8609, TK: 0000-0001-6662-4375.
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. JZ: 0000-0002-6192-4632, PAT: 0000-0002-1992-2086, EEB: 0000-0002-5959-6190.
[*§]Equal and Corresponding authors: ELS: emma.schymanski@uni.lu and EEB: bolton@ncbi.nlm.nih.gov

## Abstract

Per- and polyfluoroalkyl substances (PFAS) are of high concern, with calls to regulate these as a class. In 2021, the Organisation for Economic Co-operation and Development (OECD) revised the definition of PFAS to include any chemical containing at least one saturated $CF_2$ or $CF_3$ moiety. The consequence is that one of the largest open chemical collections, PubChem, with 115 million compounds, now contains over 7 million PFAS under this revised definition. These numbers are several orders of magnitude higher than previously established PFAS lists (typically thousands of entries) and pose an incredible challenge to researchers and computational workflows alike. This article describes a dynamic, openly accessible effort to navigate and explore the >7 million PFAS and >21 million fluorinated compounds (17 June 2023) in PubChem by establishing the “PFAS and Fluorinated Compounds in PubChem” Classification Browser (or “PubChem PFAS Tree”). A total of 36,500 nodes support browsing of the content according to several categories, including classification, structural properties, regulatory status, or presence in existing PFAS suspect lists. Additional annotation and associated data can be used to create subsets (and thus manageable suspect lists or databases) of interest for a wide range of environmental, regulatory, exposomics and other applications.

## Keywords

Per- and polyfluoroalkyl substances, chemical database, classification, chemical regulation, exposure, high resolution mass spectrometry, identification, open science

## Synopsis

The open “PFAS and Fluorinated Chemicals in PubChem” collection helps explore millions of PFAS and create relevant subsets for various applications.

# Introduction

Per- and polyfluoroalkyl substances (PFAS) are a group of substances of high environmental and toxicological concern gaining increasing attention due to bioaccumulation and a growing reputation as "forever chemicals" – so much so that there is now a drive to treat PFAS as a class for environmental regulation[1]. The 2011 definition of PFAS by Buck *et al.*[2] included substances as PFAS if they contained two (or more) connected saturated $CF_2$ groups. In 2021, the Organisation for Economic Co-operation and Development (OECD) revised the definition of PFAS in ENV/CBC/MONO(2021)25[3] as follows: "*PFAS are defined as fluorinated substances that contain* ***at least one fully fluorinated methyl or methylene carbon atom (without any H/Cl/Br/I atom attached to it)****, i.e. with a few noted exceptions, any chemical with at least a perfluorinated methyl group (–$CF_3$) or a perfluorinated methylene group (–$CF_2$–) is a PFAS.*"

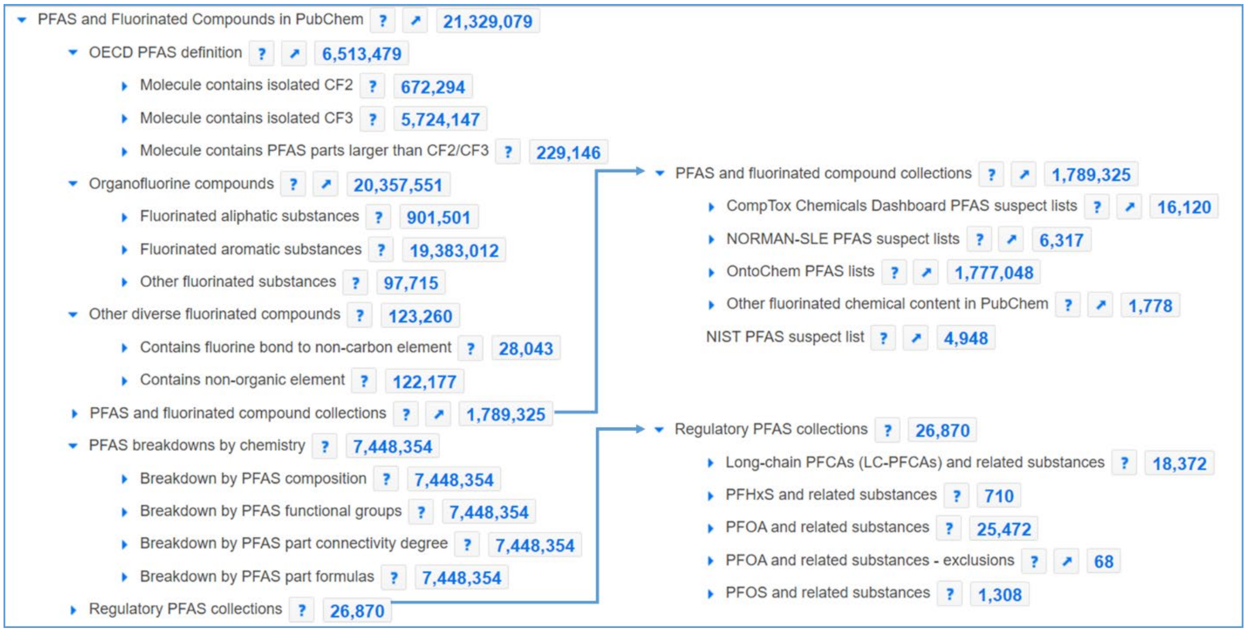
While early research efforts focused mainly on a very limited list of PFAS, the number of documented PFAS are increasing. With the emergence of high resolution mass spectrometry (HRMS) and the potential for so-called "suspect screening" for contaminants of interest using non-target analytical techniques[4,5], more extensive lists of PFAS became available. The first PFAS list hosted by the NORMAN Suspect List Exchange[6,7] (hereafter NORMAN-SLE) was the 2015 list contributed by Trier *et al.*[8], which became the basis for the OECD list of ~4700 PFAS released in 2017[9,10]. The NORMAN-SLE currently (June 2023) contains twelve PFAS lists[6,7]. The United States Environmental Protection Agency (US EPA) CompTox Chemistry Dashboard[11] also hosts chemical lists[12] and presently (June 2023) hosts 424 lists, including 51 lists matching the PFAS search term[13,14], of which 41 contain exclusively fluorinated content. The National Institute of Standards and Technology (NIST) recently coordinated a list (hereafter the "NIST PFAS Suspect List") of 4,948 entries, including expanded homologues and expert contributions[15]. Several other research efforts describe PFAS lists, with various degrees of availability. The OECD PFAS collection of ~4700 PFAS[9,10] and the US EPA PFASMASTER list (~10,000 PFAS in 2020, currently 12,034 entries in June 2023)[16] are two of the most frequently used PFAS lists in suspect screening. Both lists also contain entries that are not discrete chemicals, *i.e.*, they also include polymer and substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs)[17]. A recent effort with Google and OntoChem investigated the influence of PFAS definition on the number of PFAS extracted from literature (CORE repository) and patents (Google Patent set), resulting in PFAS lists of between 3,457 (CORE, Buck *et al.*[2] definition) and 1,783,651 (Patent set, 2021 OECD PFAS[3] definition) discrete chemicals[18]. At the time, over 200,000 of these PFAS were not in PubChem[19,20], one of the largest open chemistry databases, but were deposited soon thereafter[18].

There have been several attempts to classify and group PFAS to help answer different questions. The comprehensive OECD efforts[9,10] contained detailed classifications. The "splitPFAS" method for automated classification was developed and tested on five of these categories[21]. Recently, overviews of PFAS uses have emerged[22], while others have looked at strategies for grouping PFAS for the protection of human and environmental health[23], or narrowed the OECD PFAS list down to those of commercial relevance, estimated to be ~6 % of the total list[24]. Most, if not all, of these approaches are still largely manual.

While integrating the NORMAN-SLE content into PubChem[6], it became clear that the number of chemicals within PubChem (115 million chemicals, June 2023) that could satisfy the 2021 OECD PFAS definition dwarfed the several thousand entries in the common PFAS suspect lists. A simple substructure search for "$CF_2$" revealed millions of potential matches in PubChem. Since new PFAS are emerging very rapidly, the need for a manageable, relevant, rapidly updateable open collection of PFAS for the community is increasingly obvious. This article describes efforts to develop an interactive, open, dynamic, browsable collection of PFAS content in PubChem to serve this purpose.

## Materials and Methods

The “PFAS and Fluorinated Compounds in PubChem” collection (hereafter “PubChem PFAS Tree”) is openly available and integrated into the Classification Browser of PubChem. It is designed to support the exploration and exchange of information on PFAS and fluorinated compounds within the community. This information is compiled and assembled using several different approaches, described further in the following sections. The online collection (shown with the first two layers of nodes in Figure 1) is updated frequently and is available at https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120.



- ▾ PFAS and Fluorinated Compounds in PubChem ? ↗ 21,329,079
  - ▾ OECD PFAS definition ? ↗ 6,513,479
    - ▸ Molecule contains isolated CF2 ? 672,294
    - ▸ Molecule contains isolated CF3 ? 5,724,147
    - ▸ Molecule contains PFAS parts larger than CF2/CF3 ? 229,146
  - ▾ Organofluorine compounds ? ↗ 20,357,551
    - ▸ Fluorinated aliphatic substances ? 901,501
    - ▸ Fluorinated aromatic substances ? 19,383,012
    - ▸ Other fluorinated substances ? 97,715
  - ▾ Other diverse fluorinated compounds ? 123,260
    - ▸ Contains fluorine bond to non-carbon element ? 28,043
    - ▸ Contains non-organic element ? 122,177
  - ▸ PFAS and fluorinated compound collections ? ↗ 1,789,325
  - ▾ PFAS breakdowns by chemistry ? 7,448,354
    - ▸ Breakdown by PFAS composition ? 7,448,354
    - ▸ Breakdown by PFAS functional groups ? 7,448,354
    - ▸ Breakdown by PFAS part connectivity degree ? 7,448,354
    - ▸ Breakdown by PFAS part formulas ? 7,448,354
  - ▸ Regulatory PFAS collections ? 26,870

- ▾ PFAS and fluorinated compound collections ? ↗ 1,789,325
  - ▸ CompTox Chemicals Dashboard PFAS suspect lists ? ↗ 16,120
  - ▸ NORMAN-SLE PFAS suspect lists ? ↗ 6,317
  - ▸ OntoChem PFAS lists ? ↗ 1,777,048
  - ▸ Other fluorinated chemical content in PubChem ? ↗ 1,778
  - NIST PFAS suspect list ? ↗ 4,948

- ▾ Regulatory PFAS collections ? 26,870
  - ▸ Long-chain PFCAs (LC-PFCAs) and related substances ? 18,372
  - ▸ PFHxS and related substances ? 710
  - ▸ PFOA and related substances ? 25,472
  - ▸ PFOA and related substances - exclusions ? ↗ 68
  - ▸ PFOS and related substances ? 1,308

*Figure 1: The PFAS and Fluorinated Compounds in PubChem collection showing the six top nodes and the first layer of sub-nodes; collection available at https://pubchem.ncbi.nlm.nih.gov/classification/#hid=120. Image created 17 June 2023.*

### PFAS and Fluorinated Content in PubChem

Four sections of the PubChem PFAS Tree are collated by running custom-designed PERL scripts (available on GitLab[25]) over the entirety of PubChem on a weekly basis, since the chemical content of PubChem updates daily, and annotation content weekly. The “*OECD PFAS definition”* section contains all discrete chemicals (excluding salts and mixtures) fulfilling the 2021 OECD PFAS definition[3] quoted above (hereafter termed an “OECD PFAS”), while the “*PFAS breakdowns by chemistry”* section contains all discrete chemicals, including salts and mixtures, that are an “OECD PFAS”[3]. Figure 8 of the OECD Monograph ENV/CBC/MONO(2021)25[3] also included a breakdown of organofluorine content into several aliphatic and aromatic categories; this structure is reflected in the “*Organofluorine compounds”* section of the PubChem PFAS Tree (see Figure 1). Over 100,000 fluorinated compounds in PubChem did not fit into the categories set out in the OECD Monograph, either because fluorine was connected to non-carbon atoms or due to the presence of non-organic elements (or both). These cases were separated into the “*Other diverse fluorinated compounds*” section, which was broken down into these two subsections (see Figure 1). A more detailed description of the contents of each section and how this is constructed is contained in the PubChem PFAS Tree documentation[26].

The scripts that construct the PubChem PFAS Tree[25] run over content that is publicly available. This data is found on the PubChem FTP site[27] and via openly available active programming interfaces (APIs) such as PUG REST[28,29]. The processing takes approximately two hours to complete (processing each of the 337 structure data files, as of June 2023, in parallel) via the PubChem compute environment.

At this stage, the entire PubChem PFAS Tree is constructed across the compound space only, *i.e.*, all entries within the tree are discrete chemicals that have a PubChem Compound Identifier (CID). Thus, polymers and UVCBs are not currently a part of the PubChem PFAS Tree (see Perspectives).

## Suspect Lists and Regulatory Collections in the PubChem PFAS Tree

The remaining two major sections of the PubChem PFAS Tree are compiled in a semi-automated manner using scripts in R and are integrated into construction of the entire PubChem PFAS Tree via mapping files. All code, mapping files and associated supporting files are on the Environmental Cheminformatics (ECI) GitLab pages[30]. These sections and code build likewise on publicly available PubChem functionality, some of which was custom designed to enable the work described here, including adding new classification browser functionality to PUG REST. The final integration of this content into the PubChem PFAS Tree is programmed and run in PERL, as part of the routine described in the previous section[25].

The "*PFAS and fluorinated compound collections*" contains five major sources of suspect lists (see top right inset of Figure 1), including NORMAN-SLE[6], CompTox[11], OntoChem[18], PubChem and NIST[15]. The CompTox chemical list content is retrieved programmatically from the PubChem EPA DSSTox Classification Browser[31] (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=105) and curated manually to retain only PFAS lists, which are included in the mapping file to retrieve the respective CIDs in each list via their classification hierarchy node identifier (HNID). The files containing the CIDs for the remaining four sources are hosted on the ECI GitLab pages; the URLs for each file are contained within the mapping file used for retrieval during the PubChem PFAS Tree construction. The NORMAN-SLE subsection contains all PFAS lists within the NORMAN-SLE (currently 12); one CID list was manually adjusted to remove non-PFAS entries such as counterions. The OntoChem CID lists are broken down by the three PFAS definitions and two data sources to form 6 categories. The NIST PFAS Suspect List was downloaded and deposited to PubChem (resulting in 1,232 new CIDs, *i.e.*, new compound record entries in PubChem) and updated once all new CIDs were registered. Finally, the PubChem content was compiled by identifying several fluorinated compound sections in other classification browsers, including the MeSH, Cameo and ChEBI browsers. Since the NIDs were not always stable (especially for ChEBI), these were also added by providing fixed files via the GitLab pages. These lists and mapping files are updated as necessary under full version control in GitLab[30]; all updates appear with the next PubChem PFAS Tree update.

The final section, "*Regulatory PFAS collections*" was added upon interactions with Andreas Buser from the Federal Office of the Environment (FOEN), Switzerland (see acknowledgements) to support regulatory PFAS efforts. As shown in Figure 1, inset bottom right, regulation surrounding four cases are covered: long-chain perfluorocarboxylic acids (LC-PFCAs), perfluorohexane sulfonic acid (PFHxS), perfluorooctanoic acid (PFOA), perfluorooctane sulfonic acid (PFOS) and the related substances for all cases. The fifth section deals with exclusions from the PFOA cases, which are separated to avoid "exclusions" being added to the PFOA category totals. Each section is constructed according to definitions from regulatory efforts such as the Stockholm Convention[32], European Union (EU) Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and EU Environmental Chemicals Agency (ECHA)[33,34]. The sections include several lists published with these definitions, as well as various PubChem queries to find matching content in

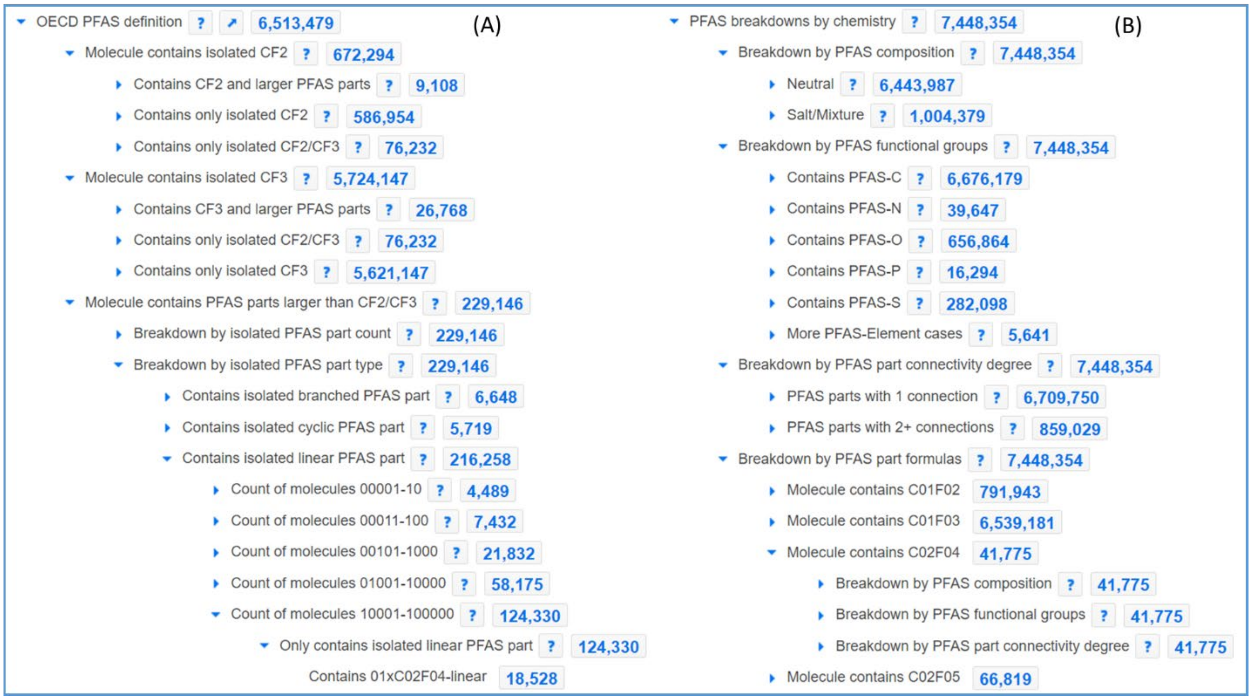
PubChem according to the definitions. Exact details of the PubChem queries are in the respective tool tips (obtained by clicking the "?" next to each heading) and in the documentation[26]. For the LC-PFCAs, the definitions came from reports UNEP/POPS/POPRC.17/7[35] and UNEP/POPS/POPRC.18/6/Add.1[36] as well as EU Regulation 2021/1297[37], with an indicative list from report UNEP/POPS/POPRC.18/INF/14[38]. For PFHxS the definitions came from UNEP/POPS/COP.10/CRP.10[39] and a draft ECHA report[40], while the initial indicative list came from UNEP/POPS/POPRC.15/INF/9[41]. The definition for PFOA came from Annex A of the Stockholm Convention (2019 revision)[32], while the initial, updated and exclusions from the PFOA lists were taken from UNEP/POPS/POPRC.17/INF/14/Rev.1[42]. Finally, the PFOS definition and PFOS listing was taken from Annex B of the Stockholm Convention[32]. The motivation and methods behind these efforts are described further in the documentation[26], as well as in a presentation at POPRC.18[43] and a webinar[44,45].

# Results and Discussion

## Overview of PFAS and Fluorinated Compounds in PubChem

As shown in Figure 1, the number of fluorinated compounds (>21 million) and PFAS (7.4 million with salts and mixtures, 6.5 million without) in PubChem is much higher than the common PFAS screening lists of four to ten thousand entries. Of the 20 million organofluorine compounds classified according to the OECD[3] (see Figure 1), ~900,000 are fluorinated aliphatic substances and 19.4 million are fluorinated aromatic substances; just under 100,000 fall into the "other" category which contain fluorine connected to non-carbon organic elements (a more detailed breakdown can be obtained by expanding the respective node in the PubChem PFAS Tree). Note that compounds can fall into more than one of these categories; the node totals always indicate the total number of CIDs under the entire node. For instance, there is no overlap between the fluorinated aliphatic and fluorinated aromatic substances, while 17,044 of the "other fluorinated substances" are also "fluorinated aromatic substances" and 7,633 are also "fluorinated aliphatic substances" (queries performed via PubChem "saved search" functionality on 17 June 2023). Approximated 120,000 fluorinated compounds fall outside the OECD organofluorine classification[3], contained within the "*Other diverse fluorinated compounds*" node.

A more detailed breakdown of the PFAS sections according to the updated OECD definition[3] is shown in Figure 2. Figure 2A reveals that 6.5 million PFAS fit this new definition (excluding salts and mixtures), of which 5.7 million contain an isolated $CF_3$ group, ~670,000 an isolated $CF_2$ group and ~230,000 a PFAS moiety larger than $CF_2$/$CF_3$ – in other words, ~230,000 PFAS also satisfy the 2011 Buck *et al.*[2] PFAS definition of substances containing at least $CF_2$-$CF_2$. As shown in Figure 2A, this can be broken down further to determine *e.g.,* how many molecules with an isolated $CF_3$ also contain larger PFAS parts (~27,000) and whether the parts are linear, branched, cyclic, and so on. As shown in the bottom of Figure 2A, the breakdown will eventually reveal the formulas of the PFAS part (here $C_2F_4$ – note the leading zeros are added to maintain a logical sorting order), should a given chain length be of interest. The total number of nodes in the tree is very high (currently 9890 nodes, June 2023). The nodes below the major sections are created dynamically depending on the data to maintain performance and functionality. As a result, formulas and other nodes appear once certain conditions are met - more details are given in the documentation[26]. Suspect lists and/or databases can be created for workflows by clicking on the nodes of interest (*i.e.*, the blue numbers), which will open a search window to either browse or download the entries. The download file contains several fields of interest; details on how to perform searches and downloads are given in the documentation[26].


(A)

- OECD PFAS definition ? ↗ 6,513,479
  - Molecule contains isolated CF2 ? 672,294
    - Contains CF2 and larger PFAS parts ? 9,108
    - Contains only isolated CF2 ? 586,954
    - Contains only isolated CF2/CF3 ? 76,232
  - Molecule contains isolated CF3 ? 5,724,147
    - Contains CF3 and larger PFAS parts ? 26,768
    - Contains only isolated CF2/CF3 ? 76,232
    - Contains only isolated CF3 ? 5,621,147
  - Molecule contains PFAS parts larger than CF2/CF3 ? 229,146
    - Breakdown by isolated PFAS part count ? 229,146
    - Breakdown by isolated PFAS part type ? 229,146
      - Contains isolated branched PFAS part ? 6,648
      - Contains isolated cyclic PFAS part ? 5,719
      - Contains isolated linear PFAS part ? 216,258
        - Count of molecules 00001-10 ? 4,489
        - Count of molecules 00011-100 ? 7,432
        - Count of molecules 00101-1000 ? 21,832
        - Count of molecules 01001-10000 ? 58,175
        - Count of molecules 10001-100000 ? 124,330
          - Only contains isolated linear PFAS part ? 124,330
            - Contains 01xC02F04-linear 18,528

(B)

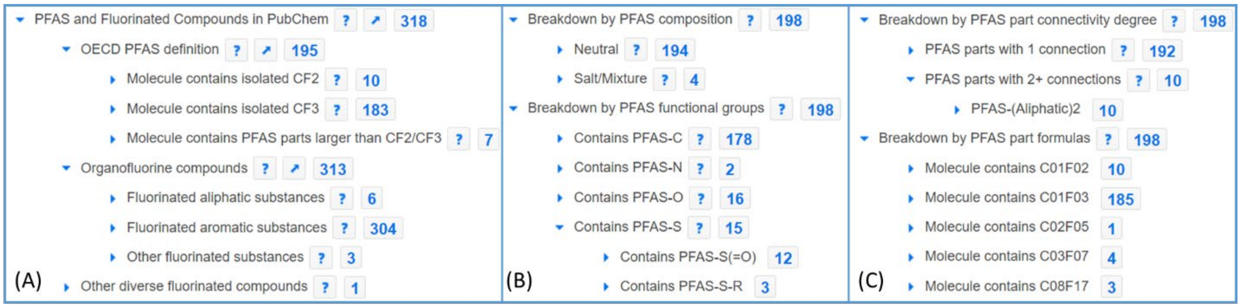
- PFAS breakdowns by chemistry ? 7,448,354
  - Breakdown by PFAS composition ? 7,448,354
    - Neutral ? 6,443,987
    - Salt/Mixture ? 1,004,379
  - Breakdown by PFAS functional groups ? 7,448,354
    - Contains PFAS-C ? 6,676,179
    - Contains PFAS-N ? 39,647
    - Contains PFAS-O ? 656,864
    - Contains PFAS-P ? 16,294
    - Contains PFAS-S ? 282,098
    - More PFAS-Element cases ? 5,641
  - Breakdown by PFAS part connectivity degree ? 7,448,354
    - PFAS parts with 1 connection ? 6,709,750
    - PFAS parts with 2+ connections ? 859,029
  - Breakdown by PFAS part formulas ? 7,448,354
    - Molecule contains C01F02 791,943
    - Molecule contains C01F03 6,539,181
    - Molecule contains C02F04 41,775
      - Breakdown by PFAS composition ? 41,775
      - Breakdown by PFAS functional groups ? 41,775
      - Breakdown by PFAS part connectivity degree ? 41,775
    - Molecule contains C02F05 66,819

*Figure 2: PFAS content according to the 2021 OECD definition. A: excluding salts and mixtures. B: including salts and mixtures. Image created 17 June 2023.*

Figure 2B shows the breakdown of PFAS including salts and mixtures, with ~1 million additional entries due to salts and mixtures. The difference in numbers on the "OECD PFAS definition" total (6.51 million) versus the "Neutral" category (6.44 million, 3rd row of Figure 2B) is due to differences in the processing as well as ambiguities in the wording of the PFAS definition. Currently, this difference is being maintained to enable an easier comparison of these "edge cases" (cyclic PFAS and PFAS-ether cases) and thus to stimulate discussion with experts within the PFAS community to help develop/refine PFAS definitions in a way that is both easy to understand and implement consistently with automated cheminformatics approaches (discussed further below). Figure 2B also reveals additional ways of browsing the PFAS content in a complementary manner to Figure 2A, including by functional groups (with the PFAS part connected to C, N, O, P, S or other elements), by connectivity (with only one connection, *i.e.*, where the PFAS is a terminal part of the molecule, or with two or more connections to the PFAS part) and by formulas, so that it is possible to search by the length of the PFAS part if a particular chain length is of interest. Again, leading zeros are present in formulas to enable a logical sort order of the formulas since the classification browser nodes appear alphabetically. The section shown in Figure 2B can be broken down by each of the respective categories, such that it is possible to exclude salts and mixtures, or only search for PFAS formulas connected to S, and so on. The dynamic "*PFAS breakdowns by chemistry*" section (Figure 2B) contains 24,600 nodes, over double the number of nodes in the "*OECD PFAS definition*" section (Figure 2A). Further details and examples are again given in the documentation[26] and explained in the webinar[44,45].

## Suspect Lists in the PubChem PFAS Tree

The suspect list section was entitled "*PFAS and fluorinated compound collections*" rather than "PFAS suspect lists" since the content of various suspect lists were not always PFAS and extremely large lists such as the OntoChem Patent collection (> 1 million entries) are too big for suspect screening. Fluorinated compounds that are not necessarily PFAS are also gaining attention as potentially harmful. For instance, there is great interest in fluorine containing pesticides and pharmaceuticals, but not all entries in the published lists (*e.g.*, lists S92[46,47] and S94[48,49] of the NORMAN-SLE, containing fluorinated pharmaceuticals[47] and pesticides[49] respectively) are PFAS. By sending these nodes to PubChem Search and subsequently Entrez, it is possible to subset the entire PubChem PFAS Tree by a given suspect list (or combination thereof) and determine which entries are PFAS, organofluorine, *etc*., as shown in Figure 3. The steps required to perform this query are explained in greater detail elsewhere[26,44,45]. The OntoChem lists, which are too big for efficient suspect screening, are already available elsewhere as database files[50]. Note that the numbers in the suspect lists in the PubChem PFAS Tree may deviate from the original lists, since only discrete chemicals are included, such that polymers and/or UVCBs will be missing (and the numbers consequently smaller) for lists containing polymer/UVCB entries in addition to discrete chemicals. Only one CompTox PFAS list (PFASMARKUSH) contained exclusively polymer/UVCB entries by design and is not displayed. While the OntoChem lists contained only discrete chemicals, these numbers also differ slightly from the published article[18] due to edge cases encountered during PubChem deposition. As discussed in Barnabas *et al.*[18], different cheminformatics toolkits perceive the structures differently: PubChem use internal code as well as the OEChem[51] and CACTVS[52] toolkits for standardization[53] and deposition to create chemical records, while OntoChem used OpenChemLib[54] to produce their final lists.



(A)

- PFAS and Fluorinated Compounds in PubChem ? ↗ 318
  - OECD PFAS definition ? ↗ 195
    - Molecule contains isolated CF2 ? 10
    - Molecule contains isolated CF3 ? 183
    - Molecule contains PFAS parts larger than CF2/CF3 ? 7
  - Organofluorine compounds ? ↗ 313
    - Fluorinated aliphatic substances ? 6
    - Fluorinated aromatic substances ? 304
    - Other fluorinated substances ? 3
  - Other diverse fluorinated compounds ? 1

(B)

- Breakdown by PFAS composition ? 198
  - Neutral ? 194
  - Salt/Mixture ? 4
- Breakdown by PFAS functional groups ? 198
  - Contains PFAS-C ? 178
  - Contains PFAS-N ? 2
  - Contains PFAS-O ? 16
  - Contains PFAS-S ? 15
    - Contains PFAS-S(=O) 12
    - Contains PFAS-S-R 3

(C)

- Breakdown by PFAS part connectivity degree ? 198
  - PFAS parts with 1 connection ? 192
  - PFAS parts with 2+ connections ? 10
    - PFAS-(Aliphatic)2 10
- Breakdown by PFAS part formulas ? 198
  - Molecule contains C01F02 10
  - Molecule contains C01F03 185
  - Molecule contains C02F05 1
  - Molecule contains C03F07 4
  - Molecule contains C08F17 3

*Figure 3: Exploring the contents of fluorine containing pesticides, S94 FLUOROPEST[48,49] with the PubChem PFAS Tree. (A) OECD PFAS (195) and organofluorine content (313 of 318 entries – the missing entries are salts). (B) Breakdown of PFAS (with salts), also showing heteroatom connections. (C) Breakdown by connectivity degree and PFAS part formulas, revealing most pesticides contain $CF_2$ or $CF_3$. Image created 17 June 2023 using PubChem Entrez functionality, explained further in the documentation[26].*

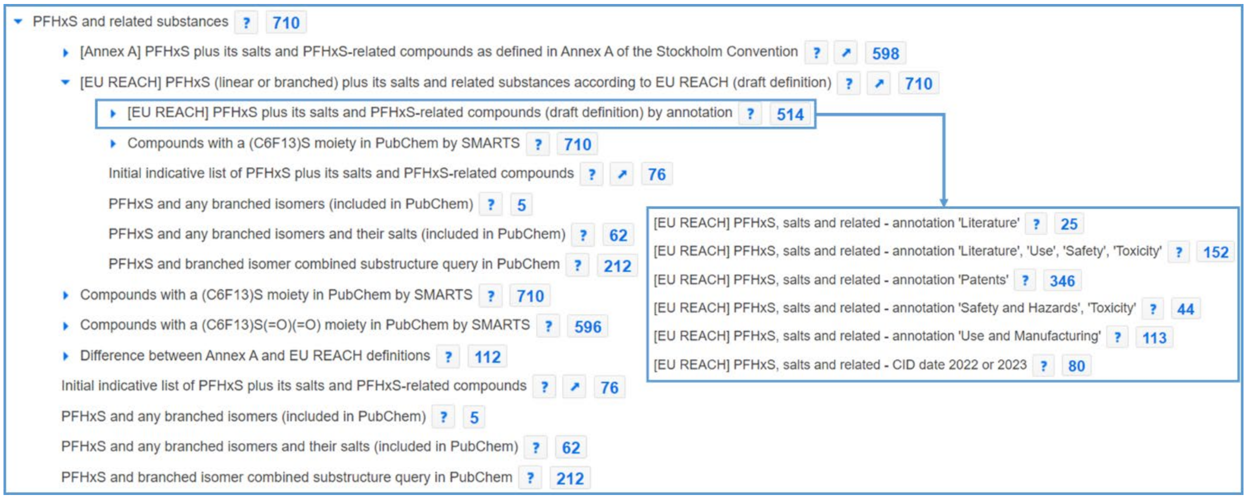
This "*PFAS and fluorinated compound collections*" section is also designed to enable the addition of new PFAS or fluorinated content into PubChem as they are documented, to fill gaps in the database and ensure rapid discovery of new and relevant entries by the community. The necessity for a rapid discovery of new PFAS of concern is one motivation for the regular updates of the entire PubChem PFAS Tree. As mentioned above, the integration of these collections has resulted in the addition of >200,000 new PFAS entries to PubChem, including >200,000 from OntoChem, 1,232 from the NIST PFAS Suspect List and several entries from both the CompTox and NORMAN-SLE contributions, which have been deposited progressively over several years. Almost 25% of the NIST PFAS list was new content to PubChem, showing the importance of hand curated expert knowledge from researchers to fill knowledge and database gaps. The NORMAN-SLE[6,7] hosts several lists, developed using templates designed together with PubChem[55,56], which can be used to add new PFAS or other compounds as soon as a reference information is available, thus providing a channel for the scientific community to add new data to the public domain. Contact details are given in the documentation[26]. Several examples of community contributions were provided in the webinar[44,45]. The information should be available under an appropriate license (*e.g.*, CC-BY[57]) to enable inclusion.

## Regulatory Collections

The final node in the PubChem PFAS Tree, "*Regulatory PFAS collections*" allows users to investigate several aspects of PFAS regulation, including the impact of different wording in definitions under consideration on the number of compounds potentially covered by the regulation. The following paragraphs cover the different cases one by one. Further details on how to perform the search queries, overlaps, downloads and other functions mentioned below can be found in the tooltips, documentation[26] and webinar[44,45].

The "*PFOS and related substances*" section is the simplest. It contains the original eight entries for "*PFOS plus salts, isomers and PFOSF*" listed in the Stockholm Convention Annex B[32] and an extended listing of all content in PubChem matching the "*PFOS plus salts, isomers and PFOSF*" definition, currently 1,304 entries in total (first node appearing in this section, which can be expanded to see the contributing subsections/categories). This 1,304 comprises PFOS and branched isomers (18), PFOS, PFOSF and salts (239) and a merged PFOS and PFOSF substructure query to find all matching mixtures (1,291 CIDs). An additional section outlines compounds that transform to PFOS (under normal conditions, *i.e.,* excluding advanced treatment transformations) that are in PubChem for information purposes, but these four entries are not included in the extended listing of "*PFOS and related substances*".

The "*PFHxS and related substances*" section contains a lot more detail than the PFOS section, as two different definitions are currently being explored for the Stockholm Convention and EU REACH. This is an interesting example where a slight change in the wording of the definition results in a difference of over 100 CIDs (chemicals) in the resulting lists. The Stockholm Convention PFHxS definition[39] defines related compounds as compounds with a $C_6F_{13}S(=O)(=O)$ moiety (596 CIDs in total), whereas the EU REACH definition[40] defines this as $C_6F_{13}S$ (710 CIDs total). Both definitions appear at the top of the PFHxS section, with content breakdowns (indicated by blue arrows, Figure 4) to show how these have been compiled.



▾ PFHxS and related substances ? 710
▸ [Annex A] PFHxS plus its salts and PFHxS-related compounds as defined in Annex A of the Stockholm Convention ? ↗ 598
▾ [EU REACH] PFHxS (linear or branched) plus its salts and related substances according to EU REACH (draft definition) ? ↗ 710
▸ [EU REACH] PFHxS plus its salts and PFHxS-related compounds (draft definition) by annotation ? 514
▸ Compounds with a (C6F13)S moiety in PubChem by SMARTS ? 710
Initial indicative list of PFHxS plus its salts and PFHxS-related compounds ? ↗ 76
PFHxS and any branched isomers (included in PubChem) ? 5
PFHxS and any branched isomers and their salts (included in PubChem) ? 62
PFHxS and branched isomer combined substructure query in PubChem ? 212
▸ Compounds with a (C6F13)S moiety in PubChem by SMARTS ? 710
▸ Compounds with a (C6F13)S(=O)(=O) moiety in PubChem by SMARTS ? 596
▸ Difference between Annex A and EU REACH definitions ? 112
Initial indicative list of PFHxS plus its salts and PFHxS-related compounds ? ↗ 76
PFHxS and any branched isomers (included in PubChem) ? 5
PFHxS and any branched isomers and their salts (included in PubChem) ? 62
PFHxS and branched isomer combined substructure query in PubChem ? 212

[EU REACH] PFHxS, salts and related - annotation 'Literature' ? 25
[EU REACH] PFHxS, salts and related - annotation 'Literature', 'Use', 'Safety', 'Toxicity' ? 152
[EU REACH] PFHxS, salts and related - annotation 'Patents' ? 346
[EU REACH] PFHxS, salts and related - annotation 'Safety and Hazards', 'Toxicity' ? 44
[EU REACH] PFHxS, salts and related - annotation 'Use and Manufacturing' ? 113
[EU REACH] PFHxS, salts and related - CID date 2022 or 2023 ? 80

*Figure 4: Regulatory collection: Example of PFHxS with two different definitions (Annex A of the Stockholm Convention and a draft definition for EU REACH). The main image shows how the main section and EU REACH sections are constructed, the inset shows the breakdown by annotation content to help find the most relevant (or recent) matches (also available for Annex A).*

For each PFHxS definition, a breakdown by major categories of annotation content has also been provided (see inset of Figure 4 for the example of EU REACH), including whether literature, use and manufacturing, safety and hazards, toxicity or patent information is available in PubChem, or whether the chemical was added only recently (CID date 2022 or 2023). In total, 598 CIDs are covered under the Stockholm Convention PFHxS definition[39], 303 with patent, 108 with use, 43 with safety or toxicity, 15 with literature information and 76 recent entries (from 2022 or 2023). The EU REACH definition contains 710 CIDs total, 346 with patent, 113 with use, 44 with safety/toxicity, 25 with literature information and 80 recent CIDs (see Figure 4 inset). The section exploring the difference between the definitions contains 112 CIDs in total, of which relatively few have either use, literature or safety/toxicity information (only 14 CIDs total).

Although PFOA, like PFOS, has been regulated already for several years, the PFOA section was much trickier to construct than the PFOS section, and remains incomplete due to the wording of the definition in Annex A of the Stockholm Convention[32]. The entire node currently contains 25,472 CIDs, but only 789 of these have been included in the "*PFOA plus its salts and PFOA-related compounds as defined in Annex A of the Stockholm Convention*" section, since the exclusions to the definition are almost impossible to define or automate cheminformatically with existing PubChem functionality. Thus, at this stage, entries that (to the best of our knowledge) meet the definition have been included, and several other sections are included under this node for users to explore other content further. The entries that are included are the selected and updated lists from the Stockholm Convention[32] (80 and 299 CIDs, respectively) plus three PubChem queries covering PFOA and branched isomers (47 CIDs), PFOA, branched isomers and salts (162 CIDs) and the PFOA plus branched isomer substructure query to capture mixtures (546 CIDs). An additional section breaks down the 789 matching PFOA content by annotation categories, such as found in literature (81), use information available (228), safety or toxicity information (41), patent information (401) or recent addition (in 2022 or 2023, 60 entries). This helps find potentially relevant entries among the hundreds of potentially regulated matches. The PFOA exclusions have been included in the node below, with placeholder nodes for content that cannot currently be created with reasonable effort. The halide exclusions have been implemented (currently 26 entries), and the updated indicative list of exclusions (35 CIDs). Polymers are inherently excluded from the tree as it currently covers compound space only, with additional functionality to enable polymer/UVCB inclusion still under active development at PubChem (and thus a potential future extension). The automatic detection of the remaining two exclusion categories, perfluoroalkyl carboxylic and phosphonic acids (including their salts, esters, halides and anhydrides) with ≥8 perfluorinated carbons, plus the perfluoroalkane sulfonic acids (including their salts, esters, halides and anhydrides) with ≥9 perfluorinated carbons has proven tricky. Although it may theoretically be possible to implement these exclusions programmatically, the current wording would require the creation of thousands of lines of custom code or several hundred very inefficient queries which, given the potentially thousands of possible matching entries, would be likewise difficult to check for accuracy and curate accordingly (several attempts at implementing this have been made already and sidelined as currently unviable). This remains an area of development for the PubChem PFAS Tree and a conversation topic with regulators, highlighting the challenges in implementing the current definition into an automated cheminformatics workflow, which will be necessary to update these regulatory lists in a manner scalable to the current numbers of PFAS (millions).

Like PFOA, the LC-PFCAs section remains difficult to complete due to the sheer number of chemicals involved. This is primarily due to the wording choice in the definition for the "related chemicals". As for PFHxS, two definitions are being explored for LC-PFCAs, the Stockholm Convention nomination of $C_9$-$C_{21}$ LC-PFCAs[35] and the EU REACH definition of $C_9$-$C_{14}$ LC-PFCAs[37]. The CIDs contained within these sections currently fulfil the LC-PFCAs, branched isomers, salts and mixture requirements of the regulation, but have not been extended to the related substances which, even in the current incomplete state, covers an additional 18,275 entries (the "related substances" sub-section remains as work in progress as the functionality required to perform these queries efficiently and automatically is still being developed). The

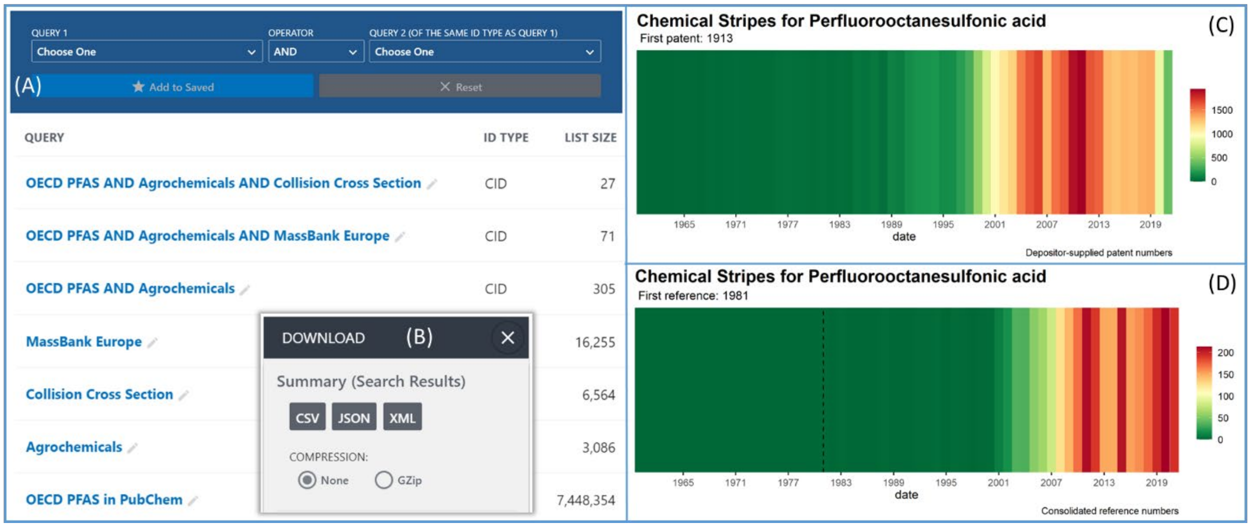
$C_9$-$C_{14}$ LC-PFCAs section is constructed using the "*PFAS breakdowns by chemistry*" section of the PubChem PFAS Tree and contains 230 CIDs. The $C_9$-$C_{21}$ LC-PFCAs section contains 745 CIDs, which includes the draft indicative listing (83 CIDs), compounds that transform to LC-PFCAs (3 CIDs), plus queries for $C_9$-$C_{21}$ LC-PFCAs, their branched isomers, salts, and mixtures. In total 584 of these have some form of annotation content, including 129 with use, 34 with safety or toxicity, 47 with literature, 490 with patent information, and finally 38 CIDs created recently (from 2022 or 2023). Again, these categories help determine which of the $C_9$-$C_{21}$ LC-PFCAs may be relevant for different use cases.

All the numbers presented in this section that have been created via PubChem queries will potentially shift with updates (most likely increasing) as the content in PubChem changes and grows.

## Interacting with the PubChem PFAS Tree

The number of PFAS contained within the PubChem PFAS Tree, let alone the number of fluorinated compounds, is overwhelming. As mentioned in previous sections, there is a large amount of data present to add context to these numbers, as well as a variety of search functions and workflows available to browse and explore the contents further to help find the most relevant PFAS or fluorinated compounds for given use cases. This section gives a brief overview of some possibilities, with further information available in the PubChem documentation[58], PubChem PFAS Tree documentation[26] and the webinars[43–45].

Every node in the PubChem PFAS Tree (*i.e.*, the blue numbers besides each category name in Figures 1-4) or any classification browser in PubChem can be sent to PubChem Search by clicking on the numbers. A separate search window will open, which allows browsing and sorting of the results, the ability to interact with individual compound records, as well as the ability to save and combine searches (see Figure 5A) or send the content to Entrez for advanced search building and/or to browse in the classification browser (see Figure 3 for example outputs). Each search query can then be downloaded in a variety of formats (see Figure 5B). It is also possible to upload custom lists to search via the PubChem landing page[20] (either pasting into the search bar, or via the "Upload ID list" option) or the PubChem Identifier Exchange[59].



| QUERY | ID TYPE | LIST SIZE |
|---|---|---|
| OECD PFAS AND Agrochemicals AND Collision Cross Section | CID | 27 |
| OECD PFAS AND Agrochemicals AND MassBank Europe | CID | 71 |
| OECD PFAS AND Agrochemicals | CID | 305 |
| MassBank Europe | | 16,255 |
| Collision Cross Section | | 6,564 |
| Agrochemicals | | 3,086 |
| OECD PFAS in PubChem | | 7,448,354 |

*Figure 5: Interacting with annotation information associated with PFAS content in PubChem. (A) The "saved searches" panel, allowing to explore the overlap (AND, OR, NOT) between various categories and searches via the blue panel at the top (query names can be edited). (B) Any search, or combination of searches, can be downloaded in a variety of formats (download fields described in the text). (C) Chemical stripes[60–62] on the patent data for PFOS, extracted from the PubChem patent tables. (D) Chemical stripes for PFOS based on consolidated reference values, extracted from PubChem consolidated references tables.*

The download file contains a number of useful fields, a selection of which will be described here (for more information see the PubChem documentation[58]). Several chemical identifier and structural information fields are included, such as names, synonyms (including CAS numbers where provided), the PubChem CID, the International Chemical Identifier (InChI)[63] and the hashed form InChIKey[63], plus the Simplified Molecular Input Line Entry System (SMILES)[64,65]. Several property fields are also given, including molecular formula, exact mass, molecular weight and octanol-water partitioning prediction (XlogP[66]). Several additional fields help add context to the chemicals, including (at the time of writing; column header in brackets) the consolidated literature count (pclidcnt), patent count (gpidcnt), annotation categories (annothits), the count of annotation (annothitcnt), the date the CID was added (cidcdate), the names of the sources who deposited this structure (sidsrcname) and the deposition categories of the sources (depcatg). The annotation categories will be discussed more in the next paragraph; note that the columns, headers, and content are potentially subject to change. The patent and literature counts have been used for many years to help prioritise chemicals in non-target identification efforts[67], but as demonstrated in Figure 5C and D, the distribution of the counts shown by the Chemical Stripes[60–62] per chemical can also reveal interesting patterns, with the patent data often increasing earlier than the literature. This means that patent data could potentially be useful to find chemicals that are being used increasingly in industry (above the trend of other chemicals) before they are discovered through problematic emissions. It is possible to find recently added CIDs using the CID date (cidcdate). Since PubChem originated in 2004, this CID date will not always be an accurate reflection of the origin date of older chemicals. For older chemicals, the literature and patent dates can help build a more accurate history, as shown in Figure 5C and D for PFOS, which was first added to PubChem in 2005, but was first mentioned in patents in 1913 and in the literature (within the collection available to PubChem) in 1981. The name of the depositors and the deposition category can help distinguish whether these chemicals come exclusively from patent literature or combinatorial libraries used for drug discovery, or whether these have been deposited by researchers, or the US EPA and so on. While these lists can be extremely long for well-known PFAS, these also tend to have substantial quantities of annotation, literature, and patent counts; the source information can help distinguish interesting entries among the long tail of matching chemicals with very little other data that potentially includes chemicals of high concern that have only just been discovered and documented.

Then annotation content of PubChem is very rich, coming from a wide variety of sources (currently over 920 data sources contribute to PubChem). The download file contains information on several major categories. The most relevant ones for environmental applications include, for example: drug and medication information; food additives and ingredients; literature; patents; pharmacology and biochemistry; safety and hazards; toxicity; use and manufacturing. The presence of these categories in the download file makes it easy to filter results by the categories of interest. Further annotation content can be browsed using the PubChem Table of Contents (TOC) Classification Browser (the “landing page” of the classification browser at https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72), which provides an overview of all annotation content in PubChem - currently 598 categories (20 June 2023). The overlap of PFAS and annotation content can be explored using the PubChem saved search and Entrez functionality. Figure 5A demonstrates how the “saved search” feature can be used to calculate how many OECD PFAS (7,448,354 CIDs, bottom row) are also agrochemicals (from the TOC heading, 3086 CIDs, of which 305 are also OECD PFAS, third row) with mass spectral data in MassBank Europe (second row: 71 CIDs that are OECD PFAS agrochemicals in MassBank Europe) or measured collision cross section (CCS) data (top row: 27 CIDs that are OECD PFAS and agrochemicals with experimental CCS values in PubChem). Each of these overlap queries can also be browsed/downloaded. Further information on how to perform these queries is available in the PubChem documentation[58], PubChem PFAS Tree documentation[26] and in the webinars[43–45].

## Perspectives

Creating a dynamic, user-friendly, browsable, and intuitive resource to explore >21 million fluorinated compounds in PubChem has been an incredibly challenging exercise in informatics and design, with several draft approaches attempted and revised before settling on the current version presented here. The functionality remains under development; automation of the regulatory and suspect list sections will be improved as the required functionality is developed. The handling of PFAS ethers ($CF_2$-O connections) and cyclic PFAS structures has been particularly challenging, along with the implementation of automated queries for the PFOA exemptions and the related compounds for the LC-PFCAs (as described above). While salts and mixtures have been added to the OECD PFAS section (resulting in an extra million CIDs included in the PubChem PFAS Tree), these are still missing in the "*Organofluorine compounds*" and "*Other diverse fluorinated compounds*" sections. With rising awareness of fluorinated counterions increasing in concentrations in wastewater and potentially becoming problematic for treatment and thus drinking water production[68], adding this is a shorter term future development, which may add a few million more CIDs to the PubChem PFAS Tree. Polymers and UVCBs will be added to the PubChem PFAS Tree once PubChem functionality is available to do so – and will likewise increase numbers further.

Community feedback has been and will continue to be valuable to help improve the design and features of future versions, potentially including the addition of new sections or substantial revision to existing sections where this is justified. The addition of the annotation content breakdowns to the regulatory collection was based on many questions from users about how to find the most relevant PFAS entries. As this annotation content is also available in the download files, it is possible to retrieve this information for any subset of the PubChem PFAS Tree. However, since the annotation data in PubChem is compiled from publicly available data and user contributions, it is not completely exhaustive. In other words, the presence of "Use and Manufacturing" information for a PFAS implies that this information is available in PubChem for that chemical with a suitable reference, but this does not imply that the entire "Use and Manufacturing" section covers all known uses.

The PubChem PFAS Tree has been available for over a year (since March 2022), was the subject of several presentations and webinars[43–45] and has already been used in published research[69]. Contributions of new PFAS or fluorinated chemicals and/or related annotation content, as well as feedback and suggestions about how the PubChem PFAS Tree can help the PFAS community answer their pressing questions are very welcome.

# Associated Content

## Supporting information

The code used to create the PubChem PFAS Tree is available on GitLab[25,30], along with more detailed documentation about the PubChem PFAS Tree[26].

# Author Information

## Author notes

*Corresponding author emails: ELS: emma.schymanski@uni.lu and EEB: bolton@ncbi.nlm.nih.gov.
§ELS and EEB contributed equally to this work.

## Author contributions

ELS: Conceptualization (equal), data curation, methodology, software, validation, writing - original draft preparation, writing - review and editing. JZ: Data curation, methodology, software. PAT: Data curation, methodology, software. PC: Data curation, validation (supporting). TK: Software. EEB: Conceptualization (equal), data curation, methodology, software (lead), validation, writing - review and editing.

## Ethics declarations

Not applicable

## Competing interests

The authors declare that they have no competing interests

## Funding sources

The work of EEB, PAT, and JZ was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health (NIH). ELS and TK acknowledge funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006, ELS and PC from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101036756, for project ZeroPM: Zero pollution of persistent, mobile substances.

# Acknowledgements

We would like to acknowledge discussions with Zhanyun Wang (EMPA, CH), Hans Peter Arp (NGI, NO), Ian Cousins, Luc Miaz, Jon Martin (ACES, SE), as well as other project members of ZeroPM. We acknowledge many valuable discussions with Andreas Buser (FOEN, CH) during development of the Regulatory collections. We would also like to acknowledge discussions and contributions from various members of the ECI and PubChem teams that were not directly involved in these efforts, including Dagny Aurich for her work on the Chemical Stripes, as well as the NORMAN Suspect List Exchange, CompTox, other resources and open science efforts in general.

# References

(1) Cousins, I. T.; DeWitt, J. C.; Glüge, J.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Scheringer, M.; Wang, Z. The High Persistence of PFAS Is Sufficient for Their Management as a Chemical Class. *Environ. Sci.: Processes Impacts* **2020**, *22* (12), 2307–2312. https://doi.org/10.1039/D0EM00355G.

(2) Buck, R. C.; Franklin, J.; Berger, U.; Conder, J. M.; Cousins, I. T.; de Voogt, P.; Jensen, A. A.; Kannan, K.; Mabury, S. A.; van Leeuwen, S. P. Perfluoroalkyl and Polyfluoroalkyl Substances in the

Environment: Terminology, Classification, and Origins. *Integrated Environmental Assessment and Management* **2011**, *7* (4), 513–541. https://doi.org/10.1002/ieam.258.

(3) OECD. *Reconciling Terminology of the Universe of Per- and Polyfluoroalkyl Substances: Recommendations and Practical Guidance*; OECD Series on Risk Management; No. 61; OECD Publishing: Paris, 2021; p 45. https://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/terminology-per-and-polyfluoroalkyl-substances.pdf (accessed 2021-11-14).

(4) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environmental Science & Technology* **2017**, *51* (20), 11505–11512. https://doi.org/10.1021/acs.est.7b02184.

(5) Liu, Y.; D'Agostino, L. A.; Qu, G.; Jiang, G.; Martin, J. W. High-Resolution Mass Spectrometry (HRMS) Methods for Nontarget Discovery and Characterization of Poly- and Per-Fluoroalkyl Substances (PFASs) in Environmental and Human Samples. *TrAC Trends in Analytical Chemistry* **2019**, *121* (115420), 115420. https://doi.org/10.1016/j.trac.2019.02.021.

(6) Mohammed Taha, H.; Aalizadeh, R.; Alygizakis, N.; Antignac, J.-P.; Arp, H. P. H.; Bade, R.; Baker, N.; Belova, L.; Bijlsma, L.; Bolton, E. E.; Brack, W.; Celma, A.; Chen, W.-L.; Cheng, T.; Chirsir, P.; Čirka, Ľ.; D'Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; Gago-Ferrero, P.; Galani, A.; Geueke, B.; Głowacka, N.; Glüge, J.; Groh, K.; Grosse, S.; Haglund, P.; Hakkinen, P. J.; Hale, S. E.; Hernandez, F.; Janssen, E. M.-L.; Jonkers, T.; Kiefer, K.; Kirchner, M.; Koschorreck, J.; Krauss, M.; Krier, J.; Lamoree, M. H.; Letzel, M.; Letzel, T.; Li, Q.; Little, J.; Liu, Y.; Lunderberg, D. M.; Martin, J. W.; McEachran, A. D.; McLean, J. A.; Meier, C.; Meijer, J.; Menger, F.; Merino, C.; Muncke, J.; Muschket, M.; Neumann, M.; Neveu, V.; Ng, K.; Oberacher, H.; O'Brien, J.; Oswald, P.; Oswaldova, M.; Picache, J. A.; Postigo, C.; Ramirez, N.; Reemtsma, T.; Renaud, J.; Rostkowski, P.; Rüdel, H.; Salek, R. M.; Samanipour, S.; Scheringer, M.; Schliebner, I.; Schulz, W.; Schulze, T.; Sengl, M.; Shoemaker, B. A.; Sims, K.; Singer, H.; Singh, R. R.; Sumarah, M.; Thiessen, P. A.; Thomas, K. V.; Torres, S.; Trier, X.; van Wezel, A. P.; Vermeulen, R. C. H.; Vlaanderen, J. J.; von der Ohe, P. C.; Wang, Z.; Williams, A. J.; Willighagen, E. L.; Wishart, D. S.; Zhang, J.; Thomaidis, N. S.; Hollender, J.; Slobodnik, J.; Schymanski, E. L. The NORMAN Suspect List Exchange (NORMAN-SLE): Facilitating European and Worldwide Collaboration on Suspect Screening in High Resolution Mass Spectrometry. *Environ Sci Eur* **2022**, *34* (1), 104. https://doi.org/10.1186/s12302-022-00680-6.

(7) NORMAN Association. *NORMAN Suspect List Exchange (NORMAN-SLE) Website*. https://www.norman-network.com/nds/SLE/ (accessed 2023-01-05).

(8) Trier, X.; Lunderberg, D. S9 | PFASTRIER | PFAS Suspect List: Fluorinated Substances. *Zenodo*, 2015, *DOI: 10.5281/zenodo.2621989*. https://doi.org/10.5281/zenodo.2621989.

(9) OECD. Toward a New Comprehensive Global Database of Per- and Polyfluoroalkyl Substances (PFASs): Summary Report on Updating the OECD 2007 List of per- and Polyfluorinated Substances (PFASs). *OECD Report* **2018**, *ENV/JM/MONO(2018)7*, 24.

(10) Wang, Z. S25 | OECDPFAS | List of PFAS from the OECD. *Zenodo*, 2018, *DOI: 10.5281/zenodo.2648776*. https://doi.org/10.5281/zenodo.2648776.

(11) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *Journal of Cheminformatics* **2017**, *9* (1), 61. https://doi.org/10.1186/s13321-017-0247-6.

(12) US Environmental Protection Agency. *CompTox Chemicals Dashboard: Chemical Lists Page*. https://comptox.epa.gov/dashboard/chemical-lists (accessed 2022-05-30).

(13) US Environmental Protection Agency. *CompTox Chemicals Dashboard: PFAS Lists*. CompTox Chemicals Dashboard: PFAS Lists. https://comptox.epa.gov/dashboard/chemical_lists/?search=PFAS (accessed 2023-06-10).

(14) Williams, A. J.; Gaines, L. G. T.; Grulke, C. M.; Lowe, C. N.; Sinclair, G. F. B.; Samano, V.; Thillainadarajah, I.; Meyer, B.; Patlewicz, G.; Richard, A. M. Assembly and Curation of Lists of Per- and Polyfluoroalkyl Substances (PFAS) to Support Environmental Science Research. *Front. Environ. Sci.* **2022**, *10*, 850019. https://doi.org/10.3389/fenvs.2022.850019.

(15) Benjamin Place. Suspect List of Possible Per- and Polyfluoroalkyl Substances (PFAS), 2021, 3 files, 1.09 MB. https://doi.org/10.18434/MDS2-2387.

(16) US EPA. *CompTox Chemicals Dashboard | PFASMASTER Chemicals*. https://comptox.epa.gov/dashboard/chemical_lists/PFASMASTER (accessed 2021-11-14).

(17) Lai, A.; Clark, A. M.; Escher, B. I.; Fernandez, M.; McEwen, L. R.; Tian, Z.; Wang, Z.; Schymanski, E. L. The Next Frontier of Environmental Unknowns: Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials (UVCBs). *Environ. Sci. Technol.* **2022**, *56* (12), 7448–7466. https://doi.org/10.1021/acs.est.2c00321.

(18) Barnabas, S. J.; Böhme, T.; Boyer, S. K.; Irmer, M.; Ruttkies, C.; Wetherbee, I.; Kondić, T.; Schymanski, E. L.; Weber, L. Extraction of Chemical Structures from Literature and Patent Documents Using Open Access Chemistry Toolkits: A Case Study with PFAS. *Digital Discovery* **2022**, *1* (4), 490–501. https://doi.org/10.1039/D2DD00019A.

(19) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Research* **2022**, gkac956. https://doi.org/10.1093/nar/gkac956.

(20) NCBI/NLM/NIH. *PubChem Website*. https://pubchem.ncbi.nlm.nih.gov/ (accessed 2023-06-10).

(21) Sha, B.; Schymanski, E. L.; Ruttkies, C.; Cousins, I. T.; Wang, Z. Exploring Open Cheminformatics Approaches for Categorizing Per- and Polyfluoroalkyl Substances (PFASs). *Environ. Sci.: Processes Impacts* **2019**, *21* (11), 1835–1851. https://doi.org/10.1039/C9EM00321E.

(22) Glüge, J.; Scheringer, M.; Cousins, I. T.; DeWitt, J. C.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Trier, X.; Wang, Z. An Overview of the Uses of Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci.: Processes Impacts* **2020**, *22* (12), 2345–2373. https://doi.org/10.1039/D0EM00291G.

(23) Cousins, I. T.; DeWitt, J. C.; Glüge, J.; Goldenman, G.; Herzke, D.; Lohmann, R.; Miller, M.; Ng, C. A.; Scheringer, M.; Vierke, L.; Wang, Z. Strategies for Grouping Per- and Polyfluoroalkyl Substances (PFAS) to Protect Human and Environmental Health. *Environ. Sci.: Processes Impacts* **2020**, *22* (7), 1444–1460. https://doi.org/10.1039/D0EM00147C.

(24) Buck, R. C.; Korzeniowski, S. H.; Laganis, E.; Adamsky, F. Identification and Classification of Commercially Relevant Per- and Poly-fluoroalkyl Substances (PFAS). *Integr Environ Assess Manag* **2021**, ieam.4450. https://doi.org/10.1002/ieam.4450.

(25) Bolton, E. E. *PubChem PFAS Tree PERL Scripts*. GitLab. https://gitlab.lcsb.uni.lu/eci/pubchem/-/tree/master/annotations/pfas/PubChem_PFAS_Tree_code (accessed 2023-06-17).

(26) Schymanski, E. L.; Chirsir, P.; Kondić, T.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. PFAS and Fluorinated Compounds in PubChem Tree: Documentation, 2023. https://gitlab.lcsb.uni.lu/eci/pubchem-docs/-/raw/main/pfas-tree/PFAS_Tree.pdf?inline=false (accessed 2023-06-10).

(27) NCBI/NLM/NIH. *PubChem Download Pages*. PubChem Download Pages. https://ftp.ncbi.nlm.nih.gov/pubchem/ (accessed 2020-05-22).

(28) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Research* **2018**, *46* (W1), W563–W570. https://doi.org/10.1093/nar/gky294.

(29) PubChem. *Programmatic Access*. https://pubchem.ncbi.nlm.nih.gov/docs/programmatic-access (accessed 2023-03-12).

(30) Schymanski, E. L.; Bolton, E. E.; Zhang, J.; Thiessen, P. A. *Environmental Cheminformatics / PubChem on GitLab: PFAS Annotations Subfolder*. GitLab. https://gitlab.lcsb.uni.lu/eci/pubchem/-/tree/master/annotations/pfas (accessed 2023-06-10).

(31) US EPA; NCBI/NLM/NIH. *PubChem Classification Browser: EPA DSSTox Tree (PubChem CompTox Chemicals Dashboard Chemical Lists Tree)*. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=105 (accessed 2022-05-30).

(32) United Nations. *Stockholm Convention on Persistent Organic Pollutants (POPs)*; Stockholm Convention on Persistent Organic Pollutants; Text and Annexes Revised in 2019 UNEP/BRS/2018/1/Rev.1; Geneva, Switzerland, 2020; p 79. https://www.pops.int/TheConvention/Overview/TextoftheConvention/tabid/2232/ (accessed 2023-06-11).

(33) European Commission. Regulation (EC) No 1907/2006 of the European Parliament and of the Council Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as Well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union* **2006**, *L 396*, 849.

(34) European Commission. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No 1907/2006. *European Commission Regulation* **2008**, *1272/2008*, 1355.

(35) United Nations. *Proposal to List Long-Chain Perfluorocarboxylic Acids, Their Salts and Related Compounds in Annexes A, B and/or C to the Stockholm Convention on Persistent Organic Pollutants*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Seventeenth meeting UNEP/POPS/POPRC.17/7; Geneva, Switzerland, 2021; p 24. https://www.pops.int/TheConvention/POPsReviewCommittee/Meetings/POPRC17/Overview/tabid/8900/Default.aspx (accessed 2023-06-10).

(36) United Nations. *Draft Risk Profile: Long-Chain Perfluorocarboxylic Acids, Their Salts and Related Compounds*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Eighteenth meeting UNEP/POPS/POPRC.18/6/Add.1*; Rome, 2022; p 56. https://www.pops.int/tabid/9165 (accessed 2023-06-10).

(37) European Commission. Commission Regulation (EU) 2021/1297 of 4 August 2021 Amending Annex XVII to Regulation (EC) No 1907/2006 of the European Parliament and of the Council as Regards Perfluorocarboxylic Acids Containing 9 to 14 Carbon Atoms in the Chain (C9-C14 PFCAs), Their Salts and C9-C14 PFCA-Related Substances. *Official Journal of the European Union* **2021**, *L 282*, 5.

(38) United Nations. *Draft Indicative List of Long-Chain Perfluorocarboxylic Acids, Their Salts and Related Compounds*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Eighteenth meeting UNEP/POPS/POPRC.18/INF/14; Rome, 2022; p 24. https://www.pops.int/tabid/9165 (accessed 2023-06-10).

(39) United Nations. *Draft Decision SC-10/[--]: Listing of Perfluorohexane Sulfonic Acid (PFHxS), Its Salts and PFHxS-Related Compounds*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Tenth meeting UNEP/POPS/COP.10/CRP.10; Geneva, Switzerland, 2021; p 1. https://www.pops.int/TheConvention/POPsReviewCommittee/Meetings/POPRC10/Overview/tabid/3779/mctl/ViewDetails/EventModID/871/EventID/514/xmid/11873/Default.aspx (accessed 2023-06-10).

(40) ECHA. *Background Document to the Opinion on the Annex XV Dossier Proposing Restrictions on Perfluorohexane Sulfonic Acid (PFHxS), Its Salts and PFHxS-Related Substances*; Draft; ECHA: Helsinki, 2020; p 304. https://echa.europa.eu/documents/10162/b4ad0be9-7a1c-e2b1-6f27-a6727c94e74b (accessed 2023-06-11).

(41) United Nations. *Initial Indicative List of Perfluorohexane Sulfonic Acid (PFHxS), Its Salts and PFHxS-Related Compounds*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Fifthteenth meeting UNEP/POPS/POPRC.15/INF/9; Rome, 2019; p 25. https://www.pops.int/TheConvention/POPsReviewCommittee/Meetings/POPRC15/Overview/tabid/8052/Default.aspx (accessed 2023-06-11).

(42) United Nations. *Updated Indicative List of Substances Covered by the Listing of Perfluorooctanoic Acid (PFOA), Its Salts and PFOA-Related Compounds*; Stockholm Convention on Persistent Organic Pollutants; Persistent Organic Pollutants Review Committee Seventeenth meeting UNEP/POPS/POPRC.17/INF/14/Rev.1; Geneva, Switzerland, 2022; p 57. https://chm.pops.int/TheConvention/POPsReviewCommittee/Meetings/POPRC17/Overview/tabid/8900/Default.aspx (accessed 2023-06-11).

(43) Schymanski, E.; Bolton, E. How Can the "PubChem PFAS Tree" Help Support the Regulation of PFAS? *POPRC.18* **2022**, *Side Event*. https://doi.org/10.5281/zenodo.7118551.

(44) Schymanski, E.; Bolton, E. ZeroPM Webinar: Are There Really 6 Million PFAS in PubChem? *Zenodo* **2023**. https://doi.org/10.5281/zenodo.7756622.

(45) *Webinar #1 Are There Really 6 Million PFAS in PubChem?  With Emma L. Schymanski and Evan E. Bolton*; ZeroPM Webinar; 2023; Vol. 2. https://www.youtube.com/watch?v=jkdvCs4pGzU (accessed 2023-06-10).

(46) Inoue, M.; Sumii, Y.; Shibata, N. S92 | FLUOROPHARMA | List of 340 ATC Classified Fluoro-Pharmaceuticals. *Zenodo*, 2022, *DOI: 10.5281/zenodo.5979647*. https://doi.org/10.5281/zenodo.5979647.

(47) Inoue, M.; Sumii, Y.; Shibata, N. Contribution of Organofluorine Compounds to Pharmaceuticals. *ACS Omega* **2020**, *5* (19), 10633–10640. https://doi.org/10.1021/acsomega.0c00830.

(48) Ogawa, Y.; Tokunaga, E.; Kobayashi, O.; Hirai, K.; Shibata, N. S94 | FLUOROPEST | List of 423 FRAC/HRAC/IRAC Classified Fluoro-Agrochemicals. *Zenodo*, 2022, *DOI: 10.5281/zenodo.6201559*. https://doi.org/10.5281/zenodo.6201559.

(49) Ogawa, Y.; Tokunaga, E.; Kobayashi, O.; Hirai, K.; Shibata, N. Current Contributions of Organofluorine Compounds to the Agrochemical Industry. *iScience* **2020**, *23* (9), 101467. https://doi.org/10.1016/j.isci.2020.101467.

(50) Barnabas, S. J.; Böhme, T.; Boyer, S.; Irmer, M.; Ruttkies, C.; Wetherbee, I.; Kondic, T.; Schymanski, E. L.; Weber, L. OntoChem PFAS CORE and Patent Files for MetFrag. *Zenodo*, 2022, *DOI: 10.5281/zenodo.6034586*. https://doi.org/10.5281/zenodo.6034586.

(51) OpenEye, Cadence Molecular Sciences. *OEChem TK | OEChem Toolkit | Cheminformatics*. https://www.eyesopen.com/oechem-tk (accessed 2023-06-17).

(52) Xemistry GmbH. *Xemistry Tools Universe*. https://www.xemistry.com/tooluniverse.shtml (accessed 2023-06-17).

(53) Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem Chemical Structure Standardization. *J Cheminform* **2018**, *10* (1), 36. https://doi.org/10.1186/s13321-018-0293-8.

(54) Actelion Pharmaceuticals Ltd. Actelion/Openchemlib, 2021. https://github.com/Actelion/openchemlib (accessed 2021-12-29).

(55) Schymanski, E. L.; Bolton, E. E. FAIR Chemical Structures in the Journal of Cheminformatics. *J Cheminform* **2021**, *13* (1), 50. https://doi.org/10.1186/s13321-021-00520-4.

(56) Schymanski, E. L.; Bolton, E. E. FAIR-Ifying the Exposome Journal: Templates for Chemical Structures and Transformations. *Exposome* **2022**, *2* (1), osab006. https://doi.org/10.1093/exposome/osab006.

(57) Schymanski, E. L.; Schymanski, S. J. Water Science Must Be Open Science. *Nat Water* **2023**, *1* (1), 4–6. https://doi.org/10.1038/s44221-022-00014-z.

(58) NCBI/NLM/NIH. *PubChem Documentation*. https://pubchem.ncbi.nlm.nih.gov/docs/about (accessed 2023-06-17).

(59) NCBI/NLM/NIH. *PubChem Identifier Exchange Service (ID Exchange)*. https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi (accessed 2022-07-23).

(60) Aurich, D. *Environmental Cheminformatics / chemicalstripes · GitLab*. GitLab. https://gitlab.lcsb.uni.lu/eci/chemicalstripes (accessed 2023-06-18).

(61) Aurich, D.; Arp, H. P.; Hale, S.; Sims, K.; Schymanski, E. Chemical Stripes – Visualizing Chemical Trends of the Past Influencing Today. *Zenodo* **2023**, *DOI:10.5281/zenodo.7885031*. https://doi.org/10.5281/zenodo.7885031.

(62) Arp, H. P. H.; Aurich, D.; Schymanski, E. L.; Sims, K.; Hale, S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. *Environ. Sci. Technol.* **2023**, *57* (16), 6355–6359. https://doi.org/10.1021/acs.est.3c01735.

(63) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *Journal of Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.

(64) Daylight Chemical Information Systems, Inc. *SMILES - A Simplified Chemical Language*. SMILES - A Simplified Chemical Language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed 2023-01-05).

(65) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

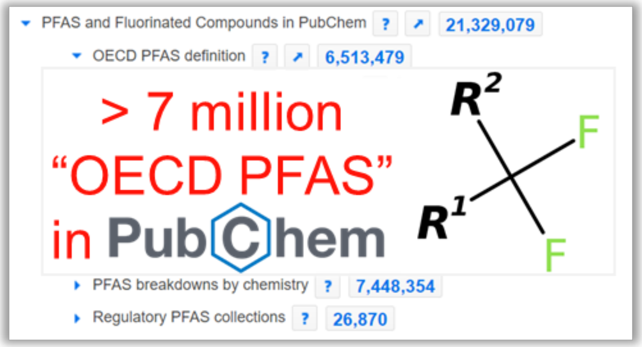
(66) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47* (6), 2140–2148. https://doi.org/10.1021/ci700257y.

(67) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation. *Journal of Cheminformatics* **2016**, *8* (1), 3. https://doi.org/10.1186/s13321-016-0115-9.

(68) Neuwald, I. J.; Muschket, M.; Seelig, A. H.; Sauter, D.; Gnirss, R.; Knepper, T. P.; Reemtsma, T.; Zahn, D. Efficacy of Activated Carbon Filtration and Ozonation to Remove Persistent and Mobile Substances – A Case Study in Two Wastewater Treatment Plants. *Science of The Total Environment* **2023**, *886*, 163921. https://doi.org/10.1016/j.scitotenv.2023.163921.

(69) Joerss, H.; Menger, F. The Complex 'PFAS World' - How Recent Discoveries and Novel Screening Tools Reinforce Existing Concerns. *Current Opinion in Green and Sustainable Chemistry* **2023**, *40*, 100775. https://doi.org/10.1016/j.cogsc.2023.100775.

# For Table of Contents Only



PFAS and Fluorinated Compounds in PubChem ? ↗ 21,329,079

OECD PFAS definition ? ↗ 6,513,479

PFAS breakdowns by chemistry ? 7,448,354

Regulatory PFAS collections ? 26,870