doi.org/10.26434/chemrxiv.12584930.v1

A Database Framework for Rapid Screening of Structure-Function Relationships in PFAS Chemistry

An Su, Krishna Rajan

Submitted date: 29/06/2020 • Posted date: 01/07/2020

Citation information: Su, An; Rajan, Krishna (2020): A Database Framework for Rapid Screening of Structure-Function Relationships in PFAS Chemistry. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12584930.v1

This paper describes a database framework that enables one to rapidly explore systematics in structure-function relationships associated with new and emerging PFAS chemistries. The data infrastructure maps high dimensional information associated with SMILES encoding of molecular structure with activity/property data. This 'PFAS-Map' serves as a 3-dimensional unsupervised visualization learning tool to automatically classify new PFAS chemistries into current well-established criteria for PFAS classification. We provide examples on how the PFAS-Map can be utilized, including the ability to predict and estimate yet unmeasured fundamental physical properties of PFAS chemistries, uncovering hierarchical characteristics in existing classification schemes and the fusion of data from diverse sources.

File list (2)

A Database Framework for Rapid Screening of Structure-Function Relationships in PFAS Chemistry

An Su* and Krishna Rajan*[+]

*Department of Materials Design and Innovation, University at Buffalo, New York, USA
[+]Corresponding author email: krajan3@buffalo.edu

ABSTRACT

This paper describes a database framework that enables one to rapidly explore systematics in structure-function relationships associated with new and emerging PFAS chemistries. The data infrastructure maps high dimensional information associated with SMILES encoding of molecular structure with activity/property data. This 'PFAS-Map' serves as a 3-dimensional unsupervised visualization learning tool to automatically classify new PFAS chemistries into current well-established criteria for PFAS classification. We provide examples on how the PFAS-Map can be utilized, including the ability to predict and estimate yet unmeasured fundamental physical properties of PFAS chemistries, uncovering hierarchical characteristics in existing classification schemes and the fusion of data from diverse sources.

INTRODUCTION

Perfluoroalkyl or polyfluoroalkyl substances (PFASs) are compounds that contain at least one fully fluorinated carbon (e.g. $-CF_3$, $-CF_2-$)[1,2]. With outstanding qualities in chemical and thermal stability, water repellency, and oil repellency, PFASs have been developed for a wide range of industrial and commercial products such as food contact materials, ski waxes, fire-fighting foams, water, and stain repellent textiles, medical devices, laboratory supplies, and personal care[1,3]. However, the presence of PFASs in freshwater systems, wildlife, and even human blood have raised serious concerns in the public domain.[4-6] The contamination and exposure of PFASs to the environment may lead to dangers with unknown consequences due to PFAS's high persistence (P), bioaccumulation potential (B), toxicity (T), and ease of being transported[7]. Although legacy PFASs such as perfluorooctanesulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) and some of their precursors are on the list for regulation or being evaluated for listing[8], alternative PFASs with similar structures and functionality as legacy PFASs, such as short-chain perfluoroalkyl carboxylic acids (PFCAs) and perfluoroalkane sulfonic acids (PFSAs), perfluoroalkyl phosphinic acids (PFPiAs), and perfluoroether carboxylic and sulfonic acids

(PFECAs and PFESAs), are still being produced and used[8-11]. Thanks to the recent developments in high-resolution mass spectrometry, increasing numbers of alternative PFASs are being discovered and analyzed, which adds thousands of structurally-identified compounds to the PFAS family[12,13]. By May 2020, there are 5,264 compounds with a defined structure under the PFAS master list provided by the United States Environmental Protection Agency (USEPA)[14].

To deal with such a large and fast-growing family of these "forever" compounds, where it is now nearly impossible to establish hazard data associated with each new PFAS chemistry, having meaningful classifications of PFAS compounds is extremely important[7,13]. A well-acknowledged PFAS classification system was published in 2011 by Buck et al with the patterns of chemical structure for each group or subgroup determined[1]. However, as more and more PFASs have been identified in the past decade, Buck's classification system needs updates in order to satisfy the need for current research. Organization for Economic Co-operation and Development (OECD) published an updated version of PFAS classification in 2018, adding new compounds to the family of PFASs such as side-chain aromatics[2]. Recent research from Wang et al. and Sha et al. have been refining the classification of PFASs as well[13,15]. As the classification of PFAS is constantly improving, an automated PFAS classification system that can reflect the updates in PFAS classification rules is needed. Machine learning approaches have found use in identifying patterns in the existing data of PFAS's properties including bioactivity, bond strength, and sources and make predictions[16-18]. However, most of the machine learning methods in these studies are supervised learning which requires the molecules' structural information as features and properties as labels, whereas the number of PFASs with known properties is significantly lower than the number of PFASs with identified structures[13]. On the other hand, unsupervised learning, an exploratory machine learning technique to find hidden patterns or grouping in data without the need of any labels[19], has not been fully utilized in the previous studies of PFASs.

In this study, we describe a database framework that maps data on structure and/or functionality and present the structure-function relationship through a 3D visualization schema (PFAS-Map). The framework takes in the identified structures of PFASs represented by Simplified Molecular Input Line Entry System (SMILES)[20] and calculates 1D and 2D molecular descriptors and PubChem fingerprints through PaDEL-descriptor[21] as the multivariate features for each compound. These high-dimensional features are reduced by principal component analysis (PCA) and the PFAS molecules are presented in a 3-dimensional PCA scores plot. In the meantime, PFASs are automatically classified in classes/subclasses based on their SMILES and molecular descriptors. The classification results, along with the data of PFAS activity/property, are also reflected in PFAS-Map. With structures, classification, and activity/property all displayed, the structure-function relationship of PFASs can be rapidly screened in PFAS-Map in an organized, straightforward way.

RESULTS

Data sources: US EPA PFAS Master List. The US EPA PFAS Master List of PFAS substances is a growing inventory that consists of various registered PFASs lists from within and outside the United States Environmental Protection Agency (US EPA), organized and structure-annotated by EPA researchers within the National Center for Computational Toxicology[14,22]. By May 2020,

the number of PFASs included in the list has increased to 7,866[14]. For our study, the chemical structures with invalid or non-canonical SMILES are removed from the list. Also, salts subgroups and isotopic specifications are removed, and ionic structures are neutralized, leaving 6,134 distinct chemical structures for further processing.

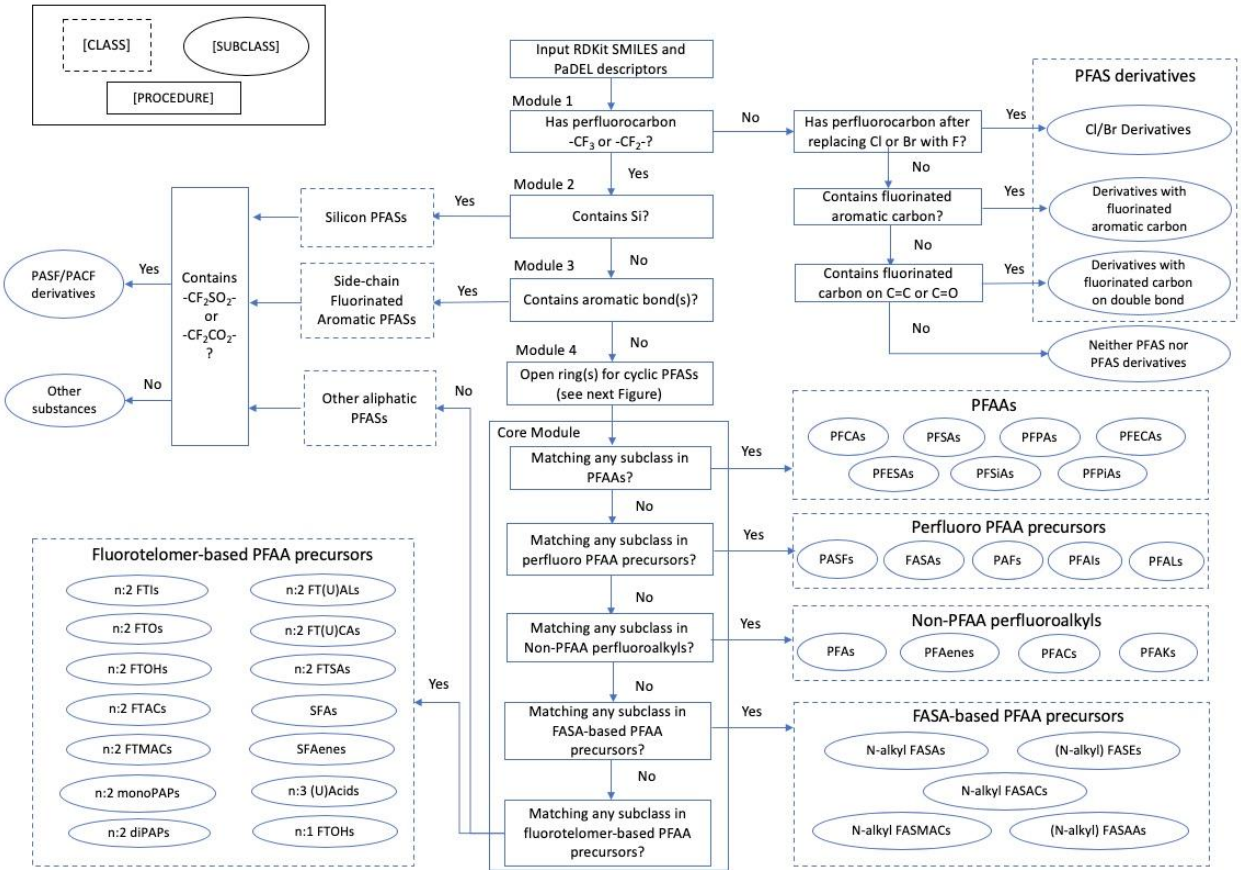

Figure 1. Structure classification of PFASs in PFAS-Map

Incorporation of Structure-Function Classification. The classification of PFAS structure consists of a series of filtering and transformation modules and a core module (Figure 1). The core modules classify the PFASs that have well-defined classes and subclasses in Buck's classification system[1] or OECD's classification[2] and its following refinements[13,15], while the filtering modules classify the rest of the PFASs (see methods for details). The incorporation of structure-function classification is implemented through the principal component analysis (PCA) which reduces ~2000 descriptors into the three dimensions that can reflect the highest variance in the structure of the PFASs. The PFASs presented as three-dimensional arrays are distributed into a 3D PCA scores plot with the structure classification results and the PFAS function data reflected as well.

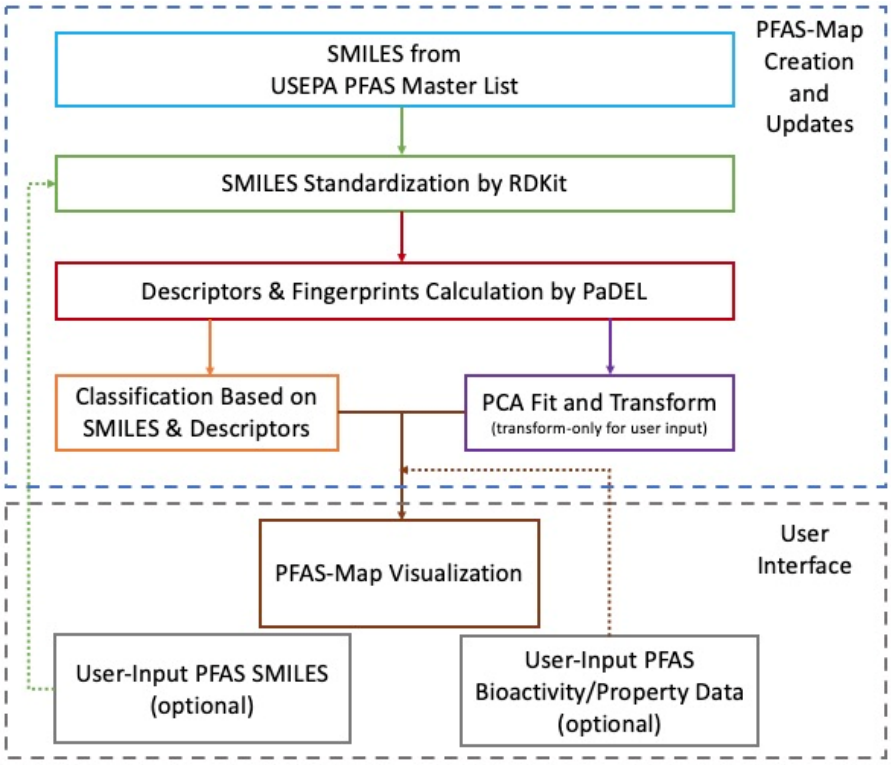

Figure 2. The architecture of PFAS-Map

Structure-Function Database Architecture. The architecture of PFAS-Map is shown in Figure 2. The key modules of PFAS-Map include SMILES standardization by RDKit[23], descriptors calculation by PaDEL[21], PFAS structure classification, PCA training and transformation, and visualization of PCA and classification results. The PFASs from US EPA PFAS Master List (EPA PFASs) are preprocessed through the framework and the output serves as the foundation of PFAS-Map. Based on this foundation, SMILES of PFASs from user input go through the same process except that the descriptors calculated are directly transformed based on the PCA model trained by EPA PFASs. Meanwhile, the user-input PFAS activity/property data can be visualized on PFAS-Map along with the PCA and classification results.

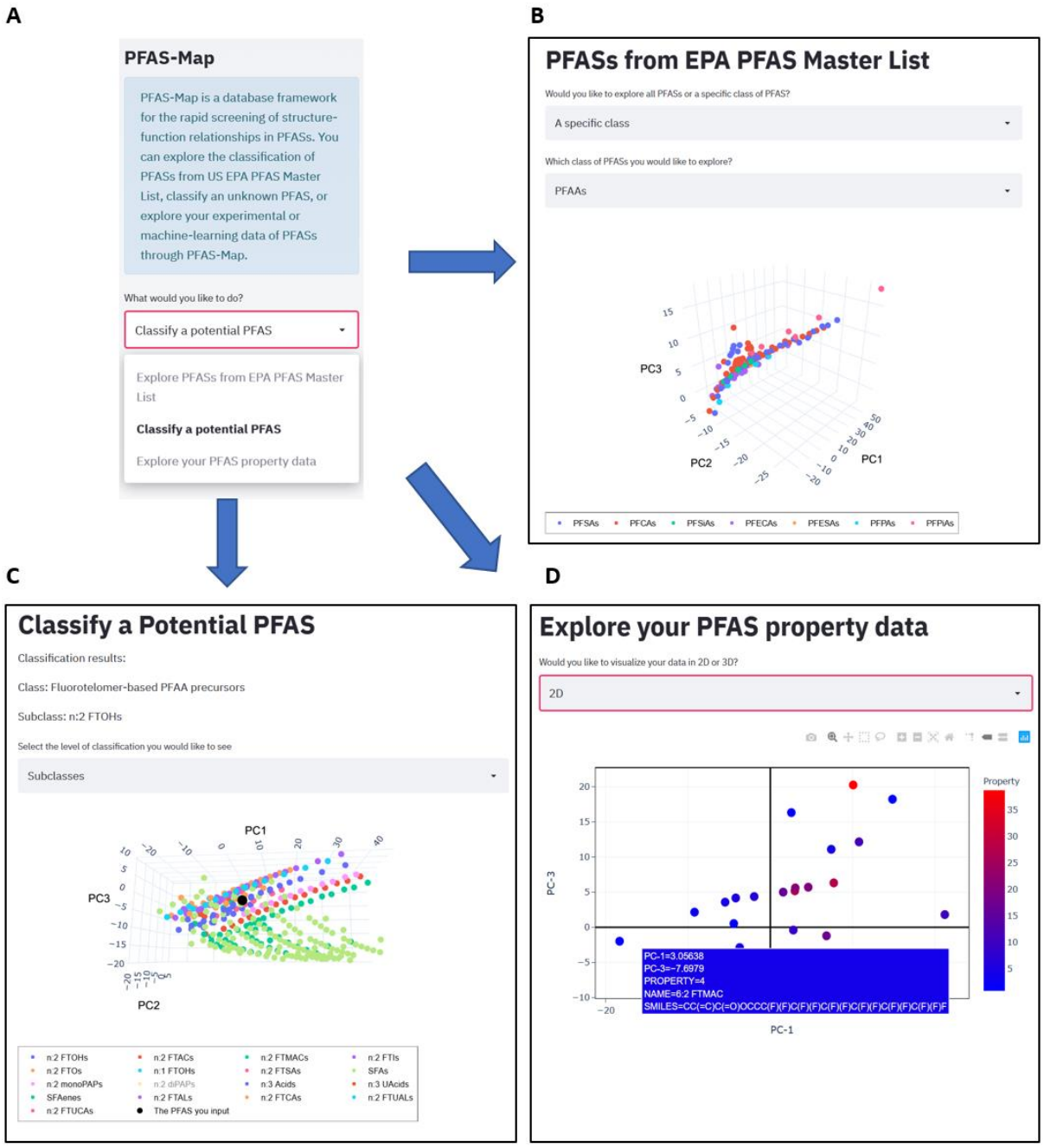

Figure 3. The user interface of PFAS-Map. A) Side bar for function selection B) Exploring EPA PFASs; C) Classifying potential PFASs; D) Explore user-input PFAS activity/property data

Some of the functionalities of PFAS-Map (Figure 3) include: the ability to query and visualize classification of PFAS chemistry in terms of molecular structure, explore similarity or dissimilarity of new or existing PFAS from the SMILES code and populate the PFAS-Map with SMILES and/or functionality information of new PFAS and readily explore and establish potentially new structure-function relationships.

DISCUSSION

In this section, we provide some examples of the utility of the PFAS-Map.

(a) Detection and visualization of sub-classifications of PFAS chemistry

Figure 4 shows a clear clustering of aromatic and aliphatic PFAS chemistries (Figure 4a) with the cluster of aromatic PFAS (light green) and aliphatic PFAS (mixed colors). In the aliphatic cluster one can observe four sub-clusters---non-PFAA perfluoroalkyls (dark green), perfluoroalkyl PFAA precursors (purple), PFAAs (dark pink), and FASA-based and fluorotelomer-based precursors (orange and blue) as is shown in Figure 4b. Hence in PFAS-Map has the capacity to capture established classifications[1,2] as well as reveal sub-classifications that would not otherwise be easily seen.

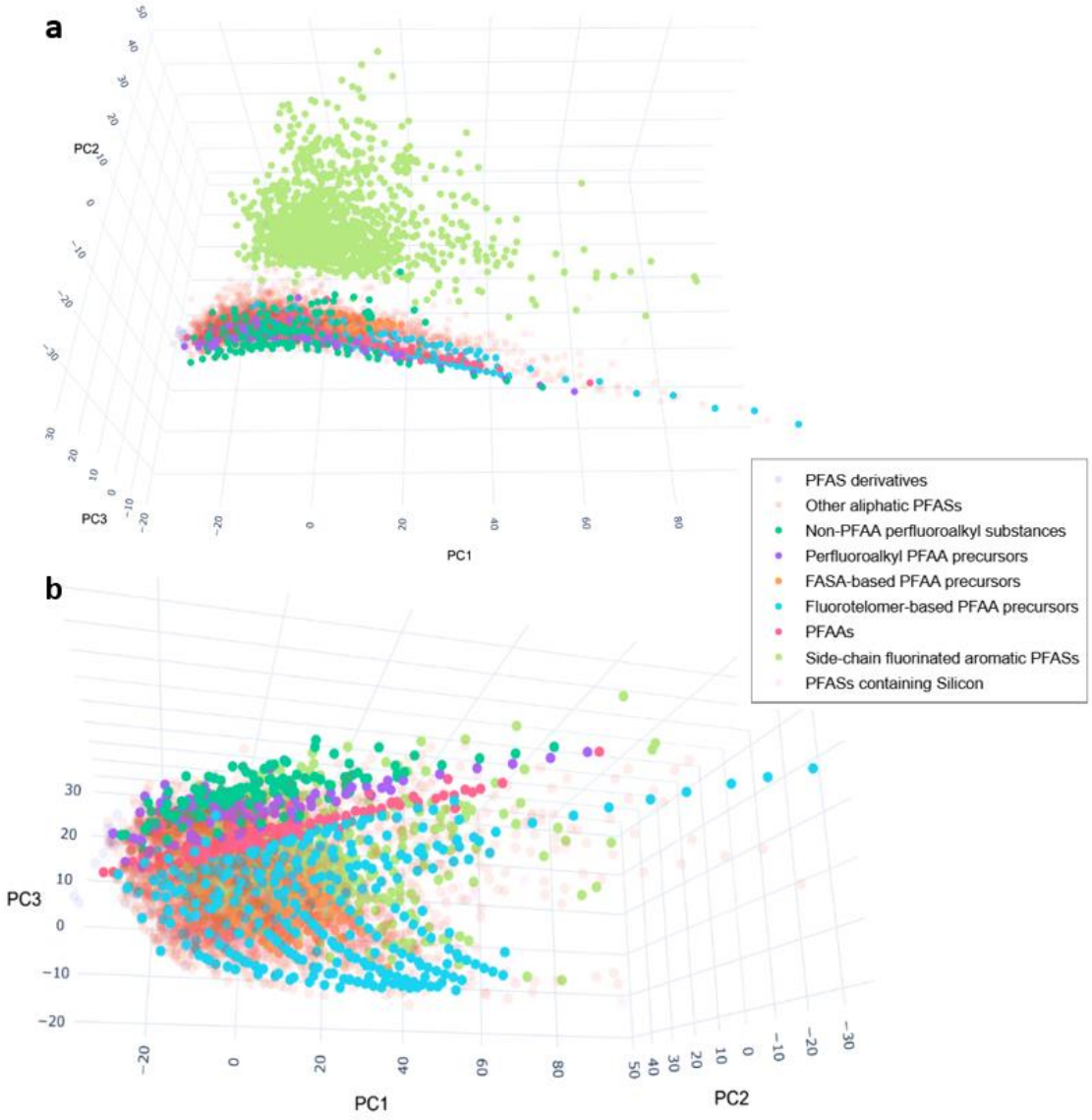

Figure 4. PFAS-Map showing EPA PFASs in classes from two different perspectives. a) The perspective showing the separation of aromatic PFASs from other PFASs. b) The perspective showing the class of aliphatic PFASs. Abbreviations: PFAA: perfluoroalkyl acids. An interactive version of this figure is provided in Supplemental Information files.

As another example, the subclasses of the two well-defined classes, FASA-based PFAA precursors and fluorotelomer-based PFAA precursors, are shown in Figure 5 and Figure 6, respectively. The subclasses in the class of FASA-based PFAA precursors follows the structural pattern as C_nF_{2n+1}-$SO_2N(C_mH_{2m})$-R^1. As viewed in the PFAS-Map, these subclasses have a clear separation from each other, and three trajectories of behavior can be tracked in the eigenspace represented in the PFAS-Map (Figure 5). First, the perfluoroalkyl chain length increases mostly as the value of PC1 increases. In addition, the sizes of N-alkyl group increase mostly as the value of PC3 decreases, which separates the compounds having the same functional group but different sizes of N-alkyl group. Furthermore, the PFASs with the same perfluoroalkyl chain but different functional groups are also separated by the PC-3 value. The n:2 fluorotelomer subclasses in the class of FASA-based PFAA precursors follows the structural pattern as C_nF_{2n+1}-C_2H_4-R^1. The distribution pattern of the n:2 fluorotelomers are similar to the FASA-based precursors---the perfluoroalkyl chain length increases mainly along +PC1 while the functional groups separate subclasses in the direction of -PC3 (Figure 6). Similar patterns in the perfluoroalkyl chain lengths, size of alkyl group(s), and the separation according to functional groups are also observed in the subclasses of other classes as is shown in the supplemental information.

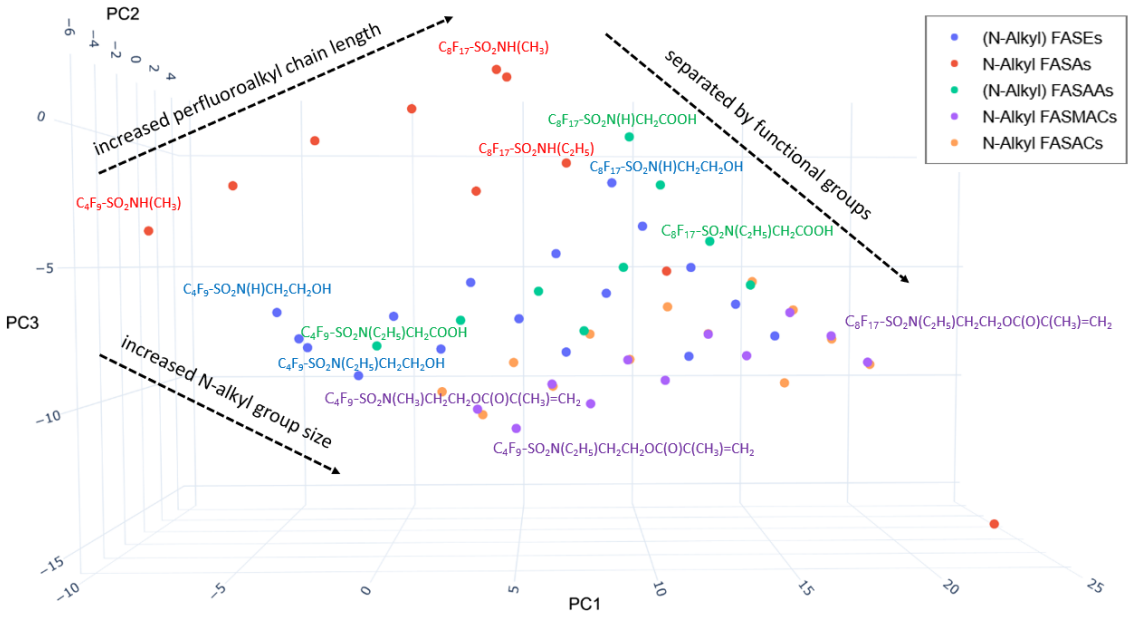

Figure 5. PFAS-Map showing all subclasses under the class of FASA-based PFAA precursors. Abbreviations: FASEs: perfluoroalkane sulfonamidoethanols; FASAs--perfluoroalkane sulfonamides; FASAAs: perfluoroalkane sulfonamidoacetic acids; FASACs: perfluoroalkane sulfonamidoethyl acrylates; FASMACs--perfluoroalkane sulfonamidoethyl methacrylates. An interactive version of this figure is provided in Supplemental Information files

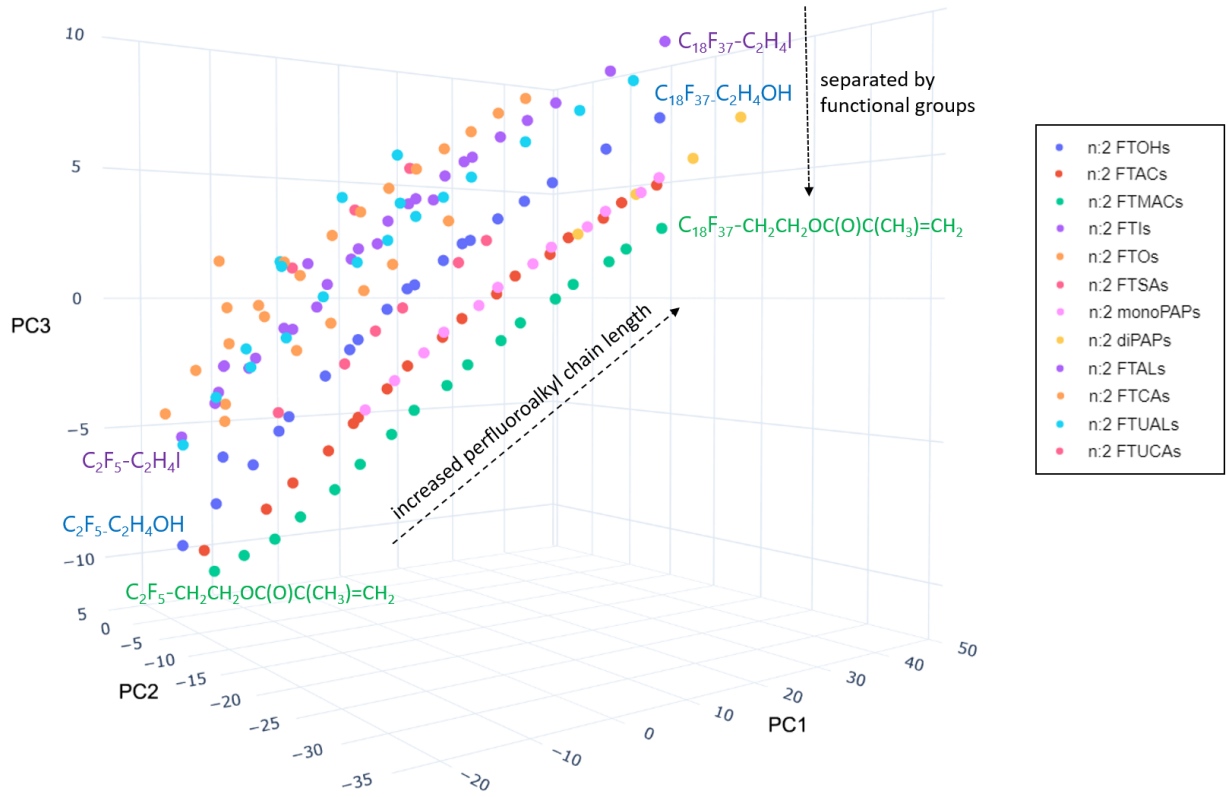

Figure 6. PFAS-Map showing n:2 fluorotelomer subclasses under the class of fluorotelomer-based PFAA precursors. Abbreviations: FTOHs: fluorotelomer alcohols; FTACs: fluorotelomer acrylates; FTMACs: fluorotelomer methacrylates; FTIs: fluorotelomer iodides; FTOs: fluorotelomer olefins; FTSAs: fluorotelomer sulfonic acids; monoPAPs: fluorotelomer phosphates, monoester; diPAPs: fluorotelomer phosphates, diester; FTALs: fluorotelomer aldehydes; FTCAs: fluorotelomer carboxylic acids; FTUALs: fluorotelomer unsaturated aldehyde; FTUCAs: fluorotelomer unsaturated carboxylic acid. An interactive version of this figure is provided in Supplemental Information files

(b)Screening the relationship between the structure and the toxicity of PFASs from two sets of experimental data.

The PFAS-Map can help one to visualize trends in experimental data on PFAS's activity/property relationships so as to uncover hidden structure-toxicity relationships that could not have been easily seen from the same data in tabular form. Weiss et al studied the competition between a series of PFASs and thyroxine (T4) for binding to the human thyroid hormone transport protein (TTR) and they showed the competition in the T4-TTR binding (%) (lower value means a higher amount of PFAS is binding to TTR)[24]. Figure 7 plots the T4-TTR binding data on the 2D projection of PFAS-Map. The binding data shows similar trends in PFCAs and PFSAs: the binding is higher (shown in red) when it comes to short chain-length (C4) or chain-length longer than C10, while the binding is the lowest (shown in blue) when it comes to C8. Hence, it is straightforward to have an estimated range of binding values for C5, C7, C9, C10, C11 for PFSAs, and C5 for PFCAs. Meanwhile, the significantly different binding values seen from the map between 2H-Perfluoro-2-octenoic acid (FTUA (6:2)) and 6:2 fluorotelomer alcohol

(FTOH (6:2)) and the high binding value for FOSAs and FOSEs infers that the T4 competition exists mostly in PFAAs but rarely in PFAA precursors.

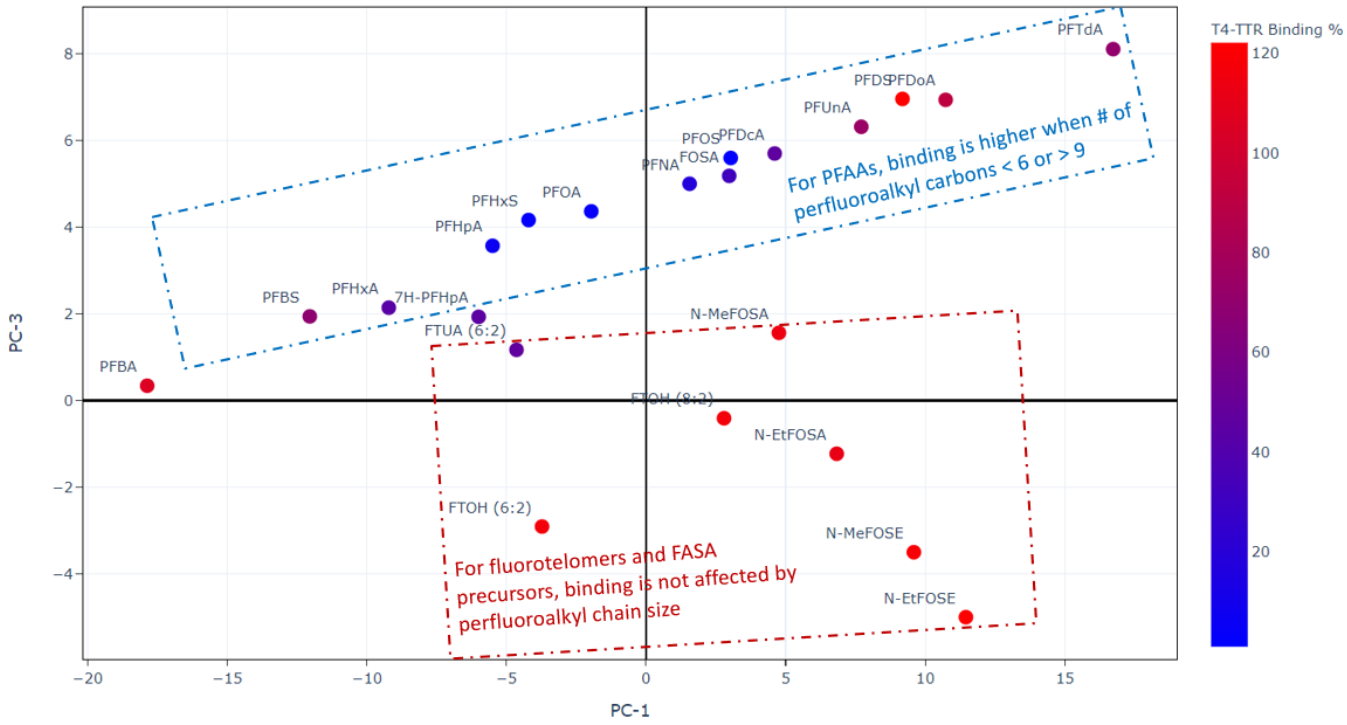

Figure 7. PFAS competed T4-TTR binding (%)[24] data shown on the 2D projection (PC1/PC3) of the PFAS-Map. Abbreviations: PFBA: perfluorobutanoic acid; PFBS: perfluorobutane sulfonic acid; PFHxA: perfluorohexanoic acid; 7H-PFHpA: 7H-perfluoroheptanoic acid; PFHpA: perfluoroheptanoic acid; PFHxS: perfluorohexane sulfonic acid; PFOA: perfluorooctanoic acid; PFNA: perfluorononanoic acid; FOSA: perfluorooctanesulfonamide; PFOS: perfluorooctanesulfonic acid; PFDcA: perfluorodecanoic acid; PFUnA: perfluoroundecanoic acid; PFDS: perfluorodecane sulfonic acid; PFDoA: perfluorododecanoic acid; PFTdA: perfluorotetradecanoic acid; FTUA (6:2): 2H-perfluoro-2-octenoic acid; N-MeFOSA: N-methylperfluorooctanesulfonamide; FTOH (8:2): 8:2 fluorotelomer alcohol; FTOH (6:2): 6:2 fluorotelomer alcohol; N-EtFOSA: N-ethylperfluorooctanesulfonamide; N-MeFOSE: N-methyl-N-(2-hydroxyethyl)perfluorooctanesulfonamide; N-EtFOSE: N-ethyl-N-(2-hydroxyethyl)perfluorooctanesulfonamide. An interactive version of this figure is provided in Supplemental Information files

The current available ToxCast data[22,25] for PFASs are visualized as in Figure 8. For each PFAS, the ToxCast number shown as the number of active tests over the number of all tests (e.g. 53/889) is converted to percentages (e.g. 6.0%) in order to be visualized in color gradients. Two significant phenomena are observed. First, most of the more active compounds are PFAAs, and an increasing active ratio is observed for PFAAs as the perfluoroalkyl chain length increases. In addition, the ToxCast number is generally lower for the non-acid PFAA precursors. By comparing the results from Figure 7 and Figure 8, we can find similarity in the structure-toxicity relationship of PFASs. For example, as one of the earliest regulated PFASs, PFOS has the most significant toxicity---it leads to one of the lowest T4-TTR protein bindings (Figure 7) and, in the meantime, has one of the highest ToxCast number (Figure 8). Also, the non-acid fluorotelomers

are generally less toxic than PFAAs based on their higher T4-TTR bindings (Figure 7) and lower ToxCast number (Figure 8), suggesting that the removal of acidic groups can possibly lower the toxicity of PFASs.

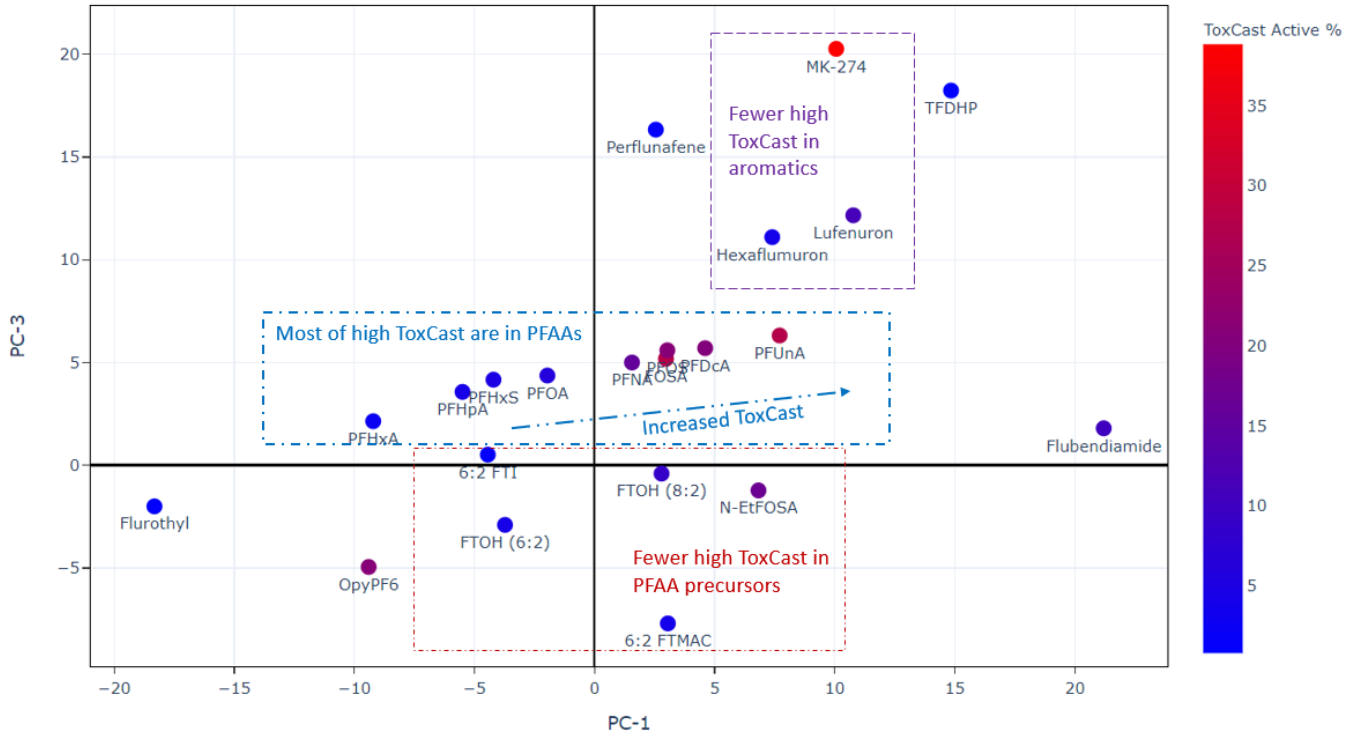

Figure 8. Currently available PFASs ToxCast data[22,25] shown on the 2D projection (PC1/PC3) of the PFAS-Map. The original ToxCast data in fraction is converted to percentage in the color bar (e.g. PFOA: 53/889 = 6.0%). Abbreviations: PFHxA: perfluorohexanoic acid; PFHpA: perfluoroheptanoic acid; PFHxS: perfluorohexane sulfonic acid; PFOA: perfluorooctanoic acid; PFNA: perfluorononanoic acid; FOSA: perfluorooctanesulfonamide; PFOS: perfluorooctanesulfonic acid; PFDcA: perfluorodecanoic acid; PFUnA: perfluoroundecanoic acid; FTOH (8:2): 8:2 fluorotelomer alcohol; FTOH (6:2): 6:2 fluorotelomer alcohol; N-EtFOSA: N-ethylperfluorooctanesulfonamide; OpyPF6: 1-methyl-3-octylimidazolium hexafluorophosphate; 6:2 FTMAC: 6:2 fluorotelomer methacrylate; 6:2 FTI: 1H,1H,2H,2H-perfluorooctyl iodide; TFDHP: tetracosafluorotetradecahydrophenanthrene. An interactive version of this figure is provided in Supplemental Information files

(c) Screening structure-activity relationships of PFAS chemicals

PFAS-Map can also be coupled with dissociation data to study the structure-persistence relationship of PFASs. Figure 9 shows the mean C-F bond dissociation energy (the average of all C-F bonds' dissociation energy in a molecule) calculated based on Raza et al.'s work on machine learning prediction of PFAS defluorination[17]. The PFAS map highlights the trend that the mean dissociation energy generally decreases as the length of perfluoroalkyl chain increases, and also that the mean dissociation energy for aromatic PFASs is significantly higher than those aliphatic PFASs with a similar number of carbons.

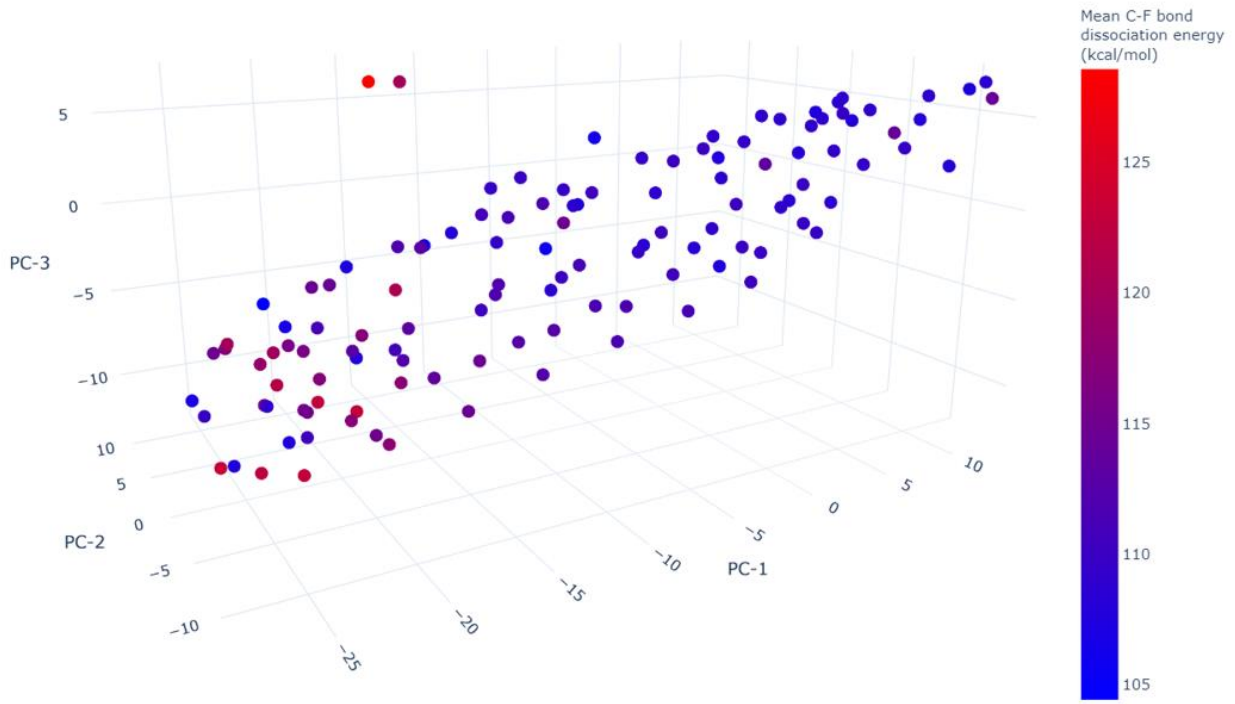

Figure 9. PFAS-Map showing the predicted mean C-F bond dissociation energy from the Raza et al.'s work "A Machine Learning Approach for Predicting Defluorination of Per and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal"[17]. An interactive version of this figure is provided in Supplemental Information files

The examples discussed above demonstrates the versatility of the PFAS Map. The automated capabilities in our database infrastructure driven by unsupervised learning methods, provides one the means to easily visualize classification patterns and trends in structures-function relationships in PFAS chemistry. One of the current bottlenecks of PFAS research is the significantly larger number of PFASs with known chemical structures compared to the number of PFASs with known properties. Hence, an unsupervised learning model like PFAS-Map fills an pressing need to appropriately classify most of PFAS molecules which are at present unlabeled in terms of their toxicity/hazard impact. The open-source nature of PFAS-Map, accommodates updates from the scientific literature on the PFAS classification rules that can be added to the source code of classification program ensuring that new classification patterns can be readily tracked.

METHODS

SMILES Standardization. The motivation for SMILES standardization is that one chemical structure can have various valid canonical SMILES generated by different computational tools or used by different databases. For example, perfluorooctanesulfonic acid (PFOS) has at least three canonical SMILES:
C(C(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(F)F)(C(C(C(F)(F)F)(F)F)(F)F)(F)F (PubChem),
OS(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F (EPA CompTox),

and O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F (RDKit). Hence, our standardization tool based on RDKit is implemented to convert SMILES from different sources into RDKit SMILES so that a RDKit-SMILES-based PFASs classification algorithm can be designed. User input SMILES goes through SMILES standardization, descriptors calculation, PFAS classification in the same way as EPA PFASs. The only difference is that the descriptors of user input PFAS will be directly transformed by the PCA model pre-trained by the EPA PFASs so that the user input PFAS and EPA PFASs can be shown in the same PFAS-Map.

Descriptors Calculation. The molecular descriptors and fingerprints of the chemical structures are calculated by PaDELPy[26], a python library for the PaDEL-descriptors software[21]. 1D and 2D molecular descriptors and PubChem fingerprints (altogether called "descriptors" in the following text) are calculated for each chemical structure. The descriptors that have invalid value for a significant number of chemical structures are removed. Simple-count descriptors (e.g. number of C, H, O, N, P, S, and F, number of aromatic atoms) are used for the classification model along with SMILES. Meanwhile, all the descriptors of EPA PFASs are used as training data for PCA.

PFAS Structure Classification. As is shown in Figure 1, module 1 filters the chemical structures not matching the most current definition of PFAS---containing "at least one -CF_3 or -CF_2- group"[1,2]. The module categorizes the unmatched chemical structures as "PFAS derivatives" if they fall into any of three subclasses: PFASs having -F substituted by -Cl or -Br, PFASs containing a fluorinated C=C carbon or C=O carbon, or PFASs containing fluorinated aromatic carbons. Otherwise, the chemical structure is marked as "not PFAS". Module 2 separates the PFASs that contain one or more Silicon atom and classify them as "Silicon PFASs" as no existing rule is available in the literature so far that can further classify the PFASs containing Silicon to our knowledge. After Module 3 filtering the side-chain fluorinated aromatics PFASs defined by OECD[2], the cyclic aliphatic PFASs are transformed to acyclic aliphatic PFASs in Module 4 by breaking the rings and add a F atom to the beginning and ending carbons of the ring. For example, O=S(=O)(O)C1(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C1(F)F (undecafluorocyclohexanesulfonic acid) is converted to O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F (perfluorohexanesulfonic acid)**.** After going through the pre-screen modules, the chemical structures that have not been categorized enter the core module of the classification system. The core module follows a "class-subclass" two-level classification, inheriting the majority of Buck's classification rules[1] for the classes including perfluoroalkyl acids (PFAAs), perfluoroalkyl PFAA precursors, perfluoroalkane-sulfonamide-based (FASA-based) PFAA precursors, and fluorotelomer-based PFAA precursors. Additional classes not in Buck's system but OECD's classification[2] and following refinements[13,15], such as perfluorinated alkanes, alkenes, alcohols, ketones, are also included as the class of non-PFAA perfluoroalkyls. In the core module, the chemical structures are tested to see if they match the structure pattern of each subclass based on their SMILES and molecular descriptors. Detailed classification algorithms can be referred in the source code.

Principal Component Analysis (PCA). A PCA model is trained with the descriptors data of EPA PFASs using Scikit-learn[27], a Python machine learning module. The trained PCA model reduced the dimensionality of the descriptors from 2090 to 3, projecting the EPA PFASs into a three-dimensional space. The trained PCA model is also used to transform the descriptors from

user-input SMILES of PFASs so that the user-input PFASs can be included in PFAS-Maps along with the EPA PFASs.

Framework Visualization. Combining the classification results with the PCA results, PFASs are visualized in a 3D interactive graph by Plotly[28] with the value of the three principal components (PCs) as the three coordinates (x, y, z) of the markers, while the colors of markers show the respective class/subclass of the PFASs. For user-input PFAS activity/property data, the data is reflected in the color of the markers or as hover text above the markers.

ACKNOWLEDGEMENTS

The authors acknowledge the support from NSF Award# 1640867 - DIBBs: EI: Data Laboratory for Materials Engineering and the Collaboratory for a Regenerative Economy (CoRE) in the Dept. of Materials Design and Innovation- University at Buffalo.

AUTHOR CONTRIBUTIONS

K.R. conceived the idea and supervised experimental work. A.S. performed the experimental studies and interpreted the results. The paper was primarily written by A.S. and edited by K. R. Both authors accepted the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The authors declare that the main data supporting the finding of this study are available within the article and its Supplementary Information files. All the supporting data have been deposited at figshare.

CODE AVAILABILITY

The code supporting the finding of this study and the code for the user interface have been deposited at figshare. Readers can follow the instructions in figshare to install the user interface at their local computer.

REFERENCES

1 Buck, R. C. *et al.* Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integrated Environmental Assessment and Management* **7**, 513-541 (2011).
2 Organisation for Economic Co-operation Development. Toward a New Comprehensive Global Database of Per-and Polyfluoroalkyl Substances (PFASs): Summary Report on Updating the OECD 2007 List of per-and Polyfluoroalkyl Substances (PFASs). (2018).
3 Cousins, I. T. *et al.* The concept of essential use for determining when uses of PFASs can be phased out. *Environmental Science: Processes & Impacts* **21**, 1803-1815 (2019).
4 Hu, X. C. *et al.* Detection of Poly- and Perfluoroalkyl Substances (PFASs) in U.S. Drinking Water Linked to Industrial Sites, Military Fire Training Areas, and Wastewater Treatment Plants. *Environmental Science & Technology Letters* **3**, 344-350 (2016).
5 Giesy, J. P. & Kannan, K. Global Distribution of Perfluorooctane Sulfonate in Wildlife. *Environmental Science & Technology* **35**, 1339-1342 (2001).
6 Hansen, K. J., Clemen, L. A., Ellefson, M. E. & Johnson, H. O. Compound-Specific, Quantitative Characterization of Organic Fluorochemicals in Biological Matrices. *Environmental Science & Technology* **35**, 766-770 (2001).
7 Ritscher, A. *et al.* Zurich Statement on Future Actions on Per- and Polyfluoroalkyl Substances (PFASs). *Environmental Health Perspectives* **126**, 084502 (2018).
8 Wang, Z., Cousins, I. T., Scheringer, M. & Hungerbuehler, K. Hazard assessment of fluorinated alternatives to long-chain perfluoroalkyl acids (PFAAs) and their precursors: Status quo, ongoing challenges and possible solutions. *Environment International* **75**, 172-179 (2015).
9 Wang, Z., Cousins, I. T., Scheringer, M. & Hungerbühler, K. Fluorinated alternatives to long-chain perfluoroalkyl carboxylic acids (PFCAs), perfluoroalkane sulfonic acids (PFSAs) and their potential precursors. *Environment International* **60**, 242-248 (2013).
10 Wang, Z., Cousins, I. T., Scheringer, M., Buck, R. C. & Hungerbühler, K. Global emission inventories for C4–C14 perfluoroalkyl carboxylic acid (PFCA) homologues from 1951 to 2030, part II: The remaining pieces of the puzzle. *Environment International* **69**, 166-176 (2014).
11 Wang, Z., Cousins, I. T., Berger, U., Hungerbühler, K. & Scheringer, M. Comparative assessment of the environmental hazards of and exposure to perfluoroalkyl phosphonic and phosphinic acids (PFPAs and PFPiAs): Current knowledge, gaps, challenges and research needs. *Environment International* **89-90**, 235-247 (2016).
12 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* **121**, 115420 (2019).
13 Wang, Z., DeWitt, J. C., Higgins, C. P. & Cousins, I. T. A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)? *Environmental Science & Technology* **51**, 2508-2518 (2017).
14 US Environmental Protection Agency. *PFAS Master List of PFAS Substances*, https://comptox.epa.gov/dashboard/chemical_lists/pfasmaster (2020).

15 Sha, B., Schymanski, E. L., Ruttkies, C., Cousins, I. T. & Wang, Z. Exploring open cheminformatics approaches for categorizing per- and polyfluoroalkyl substances (PFASs). *Environmental Science: Processes & Impacts* **21**, 1835-1851 (2019).
16 Kibbey, T. C. G., Jabrzemski, R. & O'Carroll, D. M. Supervised machine learning for source allocation of per- and polyfluoroalkyl substances (PFAS) in environmental samples. *Chemosphere* **252**, 126593 (2020).
17 Raza, A. *et al.* A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environmental Science & Technology Letters* **6**, 624-629 (2019).
18 Cheng, W. & Ng, C. A. Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List. *Environmental Science & Technology* **53**, 13970-13980 (2019).
19 Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern classification and scene analysis*. Vol. 3 (Wiley New York, 1973).
20 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31-36 (1988).
21 Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32**, 1466-1474 (2011).
22 Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics* **9**, 61-61 (2017).
23 Landrum, G. *RDKit: Open-source cheminformatics*, http://rdkit.org/ (2016).
24 Weiss, J. M. *et al.* Competitive Binding of Poly- and Perfluorinated Compounds to the Thyroid Hormone Transport Protein Transthyretin. *Toxicological Sciences* **109**, 206-216 (2009).
25 Patlewicz, G. *et al.* A Chemical Category-Based Prioritization Approach for Selecting 75 Per- and Polyfluoroalkyl Substances (PFAS) for Tiered Toxicity and Toxicokinetic Testing. *Environmental Health Perspectives* **127**, 014501 (2019).
26 Travis Kessler. *PaDELPy: A Python wrapper for PaDEL-Descriptor software*, https://github.com/ECRL/PaDELPy (2020).
27 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
28 Plotly Technologies Inc. *Collaborative data science*, https://plot.ly (2015).

Other files

Supplemental_information.zip (15.73 MiB) view on ChemRxiv • download file