MEMORANDUM

To: Eric Hooker, Kristina Hatlelid, and Charles Bevington, CPSC

From: Alexandra Maertens and Peter Ranslow, CERM
Mark Bradley and Lynne Haber, University of Cincinnati
Through: Jeanne Luh, ICF

Date: December 16, 2022

Re: Task 1: Evaluate Analog Quality for All 14 OFR Subclasses (CPSC Contract Number 61320622A0005, CPSC Order Number 61320622F2011)

---

1. Introduction

As part of the CPSC-sponsored activity that resulted in the 2019 report: *A Class Approach to Hazard Assessment of Organohalogen Flame Retardants*, the National Academies of Science, Engineering and Medicine (NASEM) subcommittee developed a list of 161 OFRs and 1,073 analogs that were categorized into 14 OFR subclasses. CPSC staff expanded the OFR inventory by considering additional data sources, resulting in a list of 488 OFRs (278 after removing mixtures and uncommonly reported congeners).

OFR chemicals that were identified by CPSC staff and not in the NASEM report currently do not have identified analogs. Additionally, an initial evaluation of NASEM analogs by CPSC staff suggests that some of the NASEM-identified analogs may not be useful because (i) they are not closely related to their matched OFR chemical(s), including physical-chemical properties, and/or (ii) there is no toxicity data available for the analog.

Under CO-1, Task 1 seeks to (i) identify analogs for the OFR chemicals identified by CPSC staff and not in the NASEM report, (ii) evaluate the quality of previously identified OFR analogs for all 14 subclasses identified by CPSC, and (iii) recommend the removal or addition of analogs from the current analog list based on analyses of chemical structure and properties and considering the availability of empirical data. The overall goal of Task 1 is to create a working set of analog chemicals for preliminary read-across analyses. To achieve this goal, Task 1 was broken out into the following three subtasks according to the project plan:

A. Identification of potentially useful analogs using two tools for all OFR substances
B. Screening of identified analogs using professional judgment and available information
C. Characterization of data availability from toxicity databases for screened analogs

This memo describes the approach taken for each of the subtasks.

2. Identification of Potentially Useful Analogs Using Two Tools for All OFR Substances (Task 1A)

To determine a list of potential analogs for the 278 CPSC expanded list of OFRs, we used PubChem and the Integrated Chemical Environment (ICE) Chemical Quest tool (referred herein as Chemquest) to identify chemicals of structural similarity. Use of two different databases allows for the creation of a more inclusive potential analog list. We describe below the approach used to search the databases and trim the list of analogs.

Tanimoto Score Choice

Both PubChem and Chemquest were searched with a minimum Tanimoto cut-off of 0.85. Initial use of a Tanimoto >0.80 produced significantly larger initial lists; use of a Tanimoto >0.90 did not result in a significant reduction in analogs or increase in potential analog quality. The choice of Tanimoto >0.85 was chosen as a reasonable cut-off to give sufficient coverage of the chemical universe without generating a large number of uninformative analogs.

For some subclasses the number of analogs was relatively few, and it may be necessary to re-search these or other databases with a lower cut-off score if data are inadequate for later stages of the project. Re-searching may be best suited to other databases, such as eChemPortal, that have more data focused searches. Using the same databases with a lower Tanimoto cutoff is unlikely to provide many additional useful analogs because of relative coverage of the chemical class and sparsity of potential analogs with measured data.

List Trimming

Several of the initial lists contained large numbers of potential analogs, with numbers in the thousands, and a selective review indicated that many of the analogs may be of low relevance or have no data (e.g. many compounds in PubChem are mentioned only as possible chemical reaction products, as opposed to potential metabolites). Several options were considered to pare the lists and preserve the largest number of high-quality analogs. We list the options considered and if applicable, the rationale for not using an option:

- Remove any chemical for PubChem without a formal identifier (e.g. IUPAC name, CAS RN) because these have low probability for measured data.
- Remove chemicals without a formal CAS RN; this was not implemented because there was concern that too many potential data sources would be missed.

We ultimately used the first option to trim our list of analogs, with the other option not implemented due to the reasons stated above. The resulting lists of potential analogs were then standardized for chemical identifiers using PubChem CIDs. Duplicates within the list were identified and eliminated where possible. In addition, some potential analogs identified were members of the subclass being queried and were removed. However, chemicals that fell within other subclasses of OFRs were not removed. Finally, analogs that were identified for more than one subclass were not removed from subsequent subclasses. Table 1 presents the number of analogs identified from each tool and the total after trimming and deduplication. Note that there

are a number of identified non-halogenated analogs in the list; because the OFRs may have non-halogenated substructures that may be relevant, those substances have been left in the initial lists at this point.

Table 1. Number of Analogs Identified by Tool, and Total Number after Trimming and Deduplication

No.	Subclass (all Polyhalogenated)	Abbreviation	PubChem Analogs	Chemquest Analogs	Total after Trimming and Deduplication
1	Organophosphates	PHOP	2,988	936	1,519
2	Bisphenol aliphatics and functionalized	PHBAF	17,002	115	7,797
3	Diphenyl ethers	PHDE	11,838	0	2,401
4	Benzene alicycles	PHBA	1,149	0	56
5	Benzene aliphatics and functionalized	PHBzAF	10,687	58	5,751
6	Phenol aliphatic ethers	PHPhAE	14,696	0	5,656
7	Alicycles	PHA	6,070	1,631	662
8	Phenol derivatives	PHPhD	1,350	0	1,016
9	Aliphatic carboxylates	PHACbx	342	21	347
10	Benzenes	PHB	21,984	279	6,590
11	Phthalates-benzoates-imides	PHPBI	5,792	0	3,916
12	Carbocycles	PHC	1,787	38	1,433
13	Triazines	PHT	129	0	111
14	Aliphatic chains	PHACh	1,765	35	1,185

Data Comparison

PubChem produced significantly greater numbers of potential analogs than Chemquest. Chemquest searches a smaller library of chemicals and uses a slightly different search algorithm for defining similarity. Specifically, Chemquest uses a SAAGAR Fingerprint Feature approach and PubChem uses a standard fingerprint algorithm. In addition, PubChem contains a data set of 299,000,000 structures; the total data set for Chemquest has not been published but is estimated to be on the order of 10,000 structures. There was, however, significant overlap between the chemicals identified from PubChem and Chemquest.

Results from Task 1A

We identified initial lists of potential analogs for all 14 subclasses based on the use of the tools and methods described above. These lists were used for database downloads in Task 1C and were further trimmed based on data availability as described below in the following sections. The actual lists of new analogs by subclass are provided as part of the evidence maps deliverable for Task 1C.

3. Screening of Identified Analogs using Professional Judgment and Available Information (Task 1B)

Work has been ongoing to review the OFR classes and analog lists in order to assess fitness for achieving project goals, potential groupings/subgroupings, and strategies to for developing read-across analysis to meet data gaps. Some of the initial work, including identification of the analogs that are also OFRs, potential clustering of the analogs, and other methods to group/evaluate the analogs for use in meeting data gaps are described below.

Clustering of Original OFR Subclasses

To explore the similarity within the subclasses, we looked at groupings of the OFR subclass chemicals by Tanimoto score. The results show significant overlap in structural features among the subclasses. The majority of the substances group into three primary clusters (Figure 1A), with multiple smaller clusters and outliers also mapped (Figure 1B). Chemicals in the smaller clusters and outliers will likely be more dependent on analog availability to make a confident judgment about toxicity vs the larger clusters. Notably, the organophosphates clustered separately from the non-organophosphates, and they can likely be best analyzed by the subclusters presented here. In most cases, the best analog for an OFR will be another OFR, although not always an OFR in the same class. Additionally, there will be limited usefulness, if any, of finding additional analogs for the chemicals within the large cluster. Even at a Tanimoto of 0.95, they largely clustered together (Figure 2), although subdividing the clusters by linkage number did show subclusters with slightly different maximum common substructures. The most common substructure within this cluster was found to be a diphenyl ether ([C]1=[C][C]=C(OC2=CC=CC=C2)[C]=[C]1] in SMILES notation) (i.e. diphenyl ether). This is not surprising, given that the polyhalogenated diphenyl ethers is the largest OFR subclass (58 members). However, it is not clear at this stage if this reflects a meaningful difference toxicologically.

Identified Analogs

The analog lists from Task 1A varied in size from 56 to nearly 8,000 per subclass. Most lists had >1,000 analogs, although the number of analogs for which toxicity data were downloaded in Task 1C was generally much smaller. The data rich compounds tended to be other OFRs already in the data set, but not necessarily in the same subclass.

Screening Based on Physicochemical Properties

In terms of molecular weight, log Kow, and Topological Polar Surface Area (TPSA), the distribution of the analogs generally matched the distribution of the original OFRs. In general, there was a narrow range of logKow and molecular weight with some variance in terms of percent halogenation – the latter is likely going to be the most useful chemoinformatic metric to further subdivide the diphenyl ether class and match analogs. Going forward, large molecular weight compounds (defined here as an MW over 900) should be treated separately from the rest of the class, and analogs for high molecular weight compounds should be matched up

based on molecular weight primarily and highest scoring Tanimoto secondarily. All classes had at least one large MW compound.

Clustering of Analogs

Similar to the clustering of the original OFRs, the bulk of the largest cluster of analogs again consisted of diphenyl ethers, which was also the class with the largest number of analogs (data not shown). However, since this subclass largely clusters together, the usefulness of additional analogs for these chemicals is difficult to ascertain unless there are significant and unexpected data gaps in the original group of OFRs. It will likely be more important to match toxicity to structure within this class to identify potential activity cliffs; if such cliffs are identified, the analogs with data may be useful on a weight of evidence (WoE) level to ascertain the validity of such structure-activity relationships.

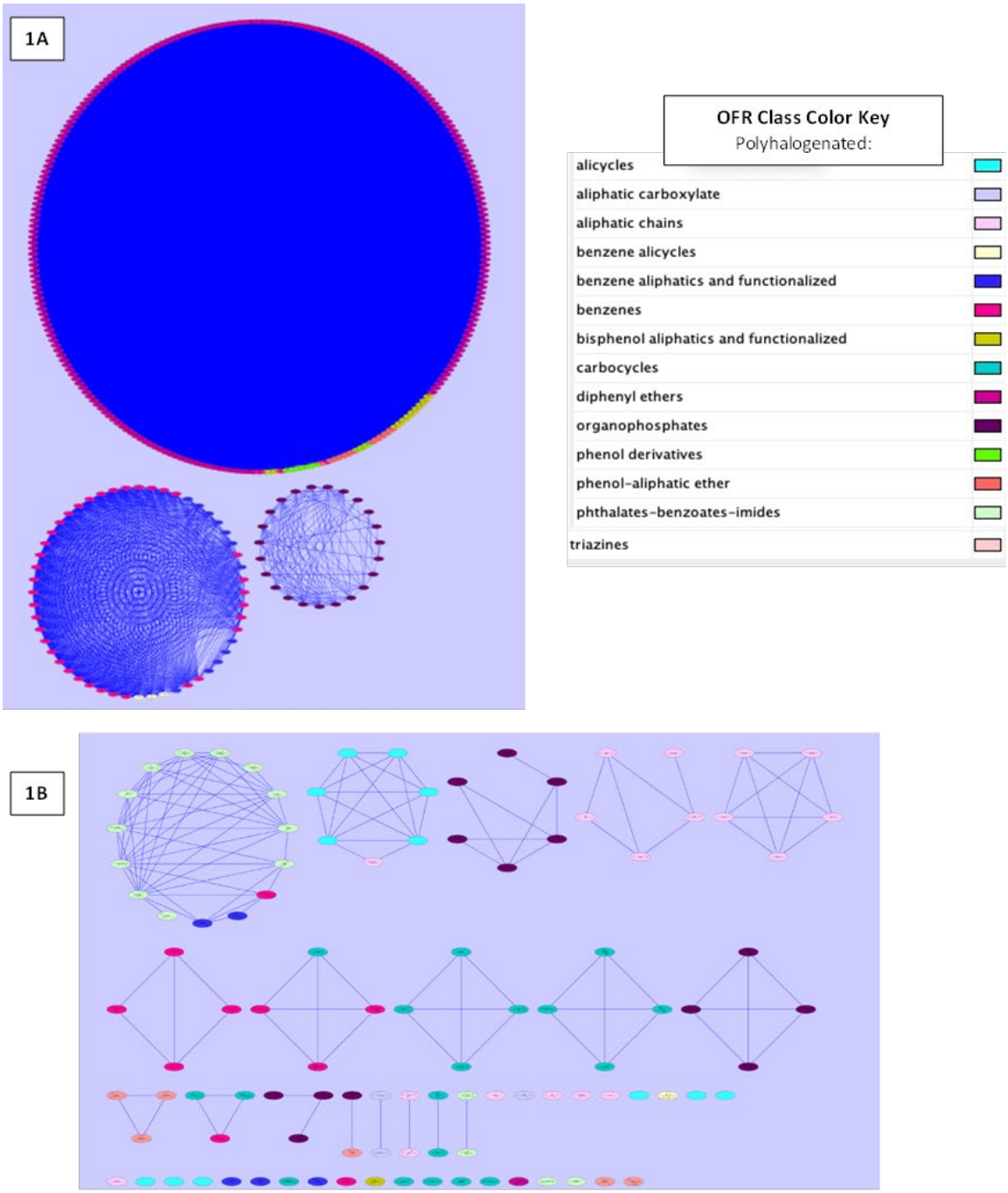

Figure 1 - Clustering analysis of OFR subclasses based on Tanimoto of 0.80; colored by class. Most chemicals were in three main clusters representing compounds across multiple subclasses (Fig 1A); this was not changed significantly even when the Tanimoto was raised to 0.85 (data not shown). While some classes clustered together, all but one organophosphate clustered with other organophosphates. Other subclass members formed smaller clusters, with a few outliers. These also grouped across multiple subclasses (Fig 1B). Chemicals without an OFR analog in this map will be more likely to be dependent on non-OFR analog data, while chemicals in the large cluster are likely to have more useful OFR analogs. The classes are color coded according to the key above.

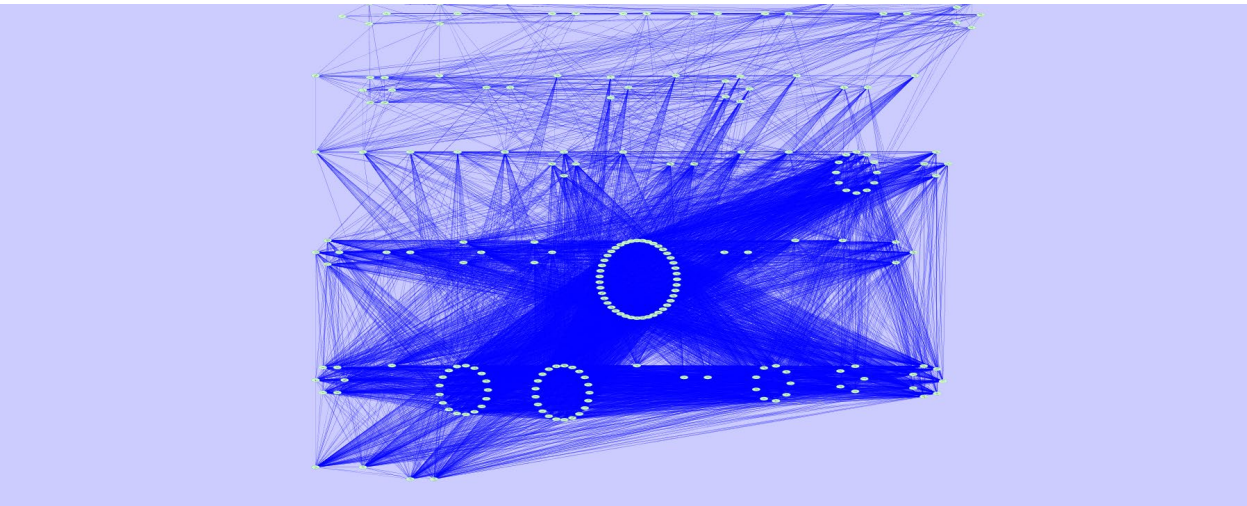

Figure 2 - Large cluster showing only linkages with Tanimoto > 0.95, sub-clustered by number of links. Maximum common substructure within this cluster, [C]1=[C][C]=C(OC2=CC=CC=C2)[C]=[C]1 (i.e. diphenyl ether), also accounted for most of the analogs; smaller clusters captured differences in halogenation. Further subdividing this cluster based on other distinguishing features, e.g. percent halogen, may be possible if structural motifs can be tied to toxicity.

Screening Based on Bioavailability

For all classes, there was a mix of predicted absorption, distribution, metabolism excretion (ADME) scores for the original OFRs (see Figure 3A and 3B) for both the potential to cross the blood-brain (BB) barrier and gastrointestinal (GI) availability. Predicted ADME is a function of molecular weight, polarity, lipophilicity, and solubility. Structurally similar chemicals often, but not always, share similar predicted ADME. Going forward, matching analogs by predicted ADME is likely the most useful chemoinformatic way of ensuring the suitability of analogs. Each subclass and analog can therefore be divided into four potential classes: low predicted bioavailability; GI availability only (i.e., is absorbed from the GI tract but does not cross the BB barrier); BB bioavailability only; and GI and BB bioavailability (see Figure 4A and 4B). Reviewing known ADME, where available, will help ground this. While bioavailability is a good first step for determining toxicity, toxicity cannot be ruled in or ruled out based on this. Nonetheless, it is critical that analogs match in terms of predicted bioavailability.

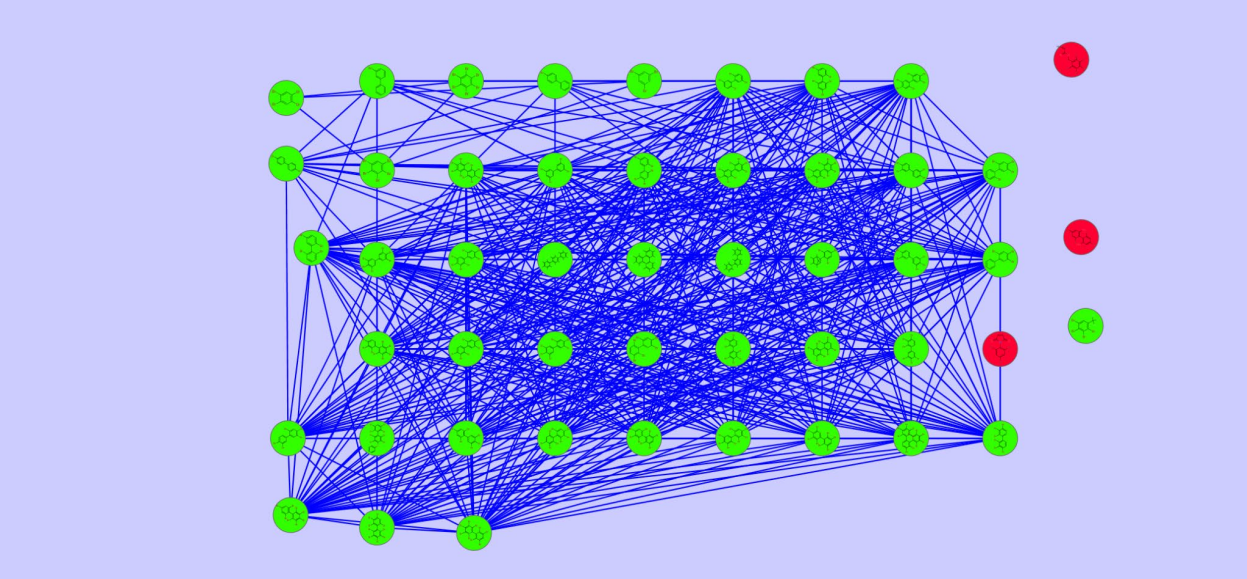

Figure 3A – All chemicals in the polyhalogenated benzene subclass; linked chemicals have a Tanimoto of 0.85; red indicates predicted high GI bioavailability.

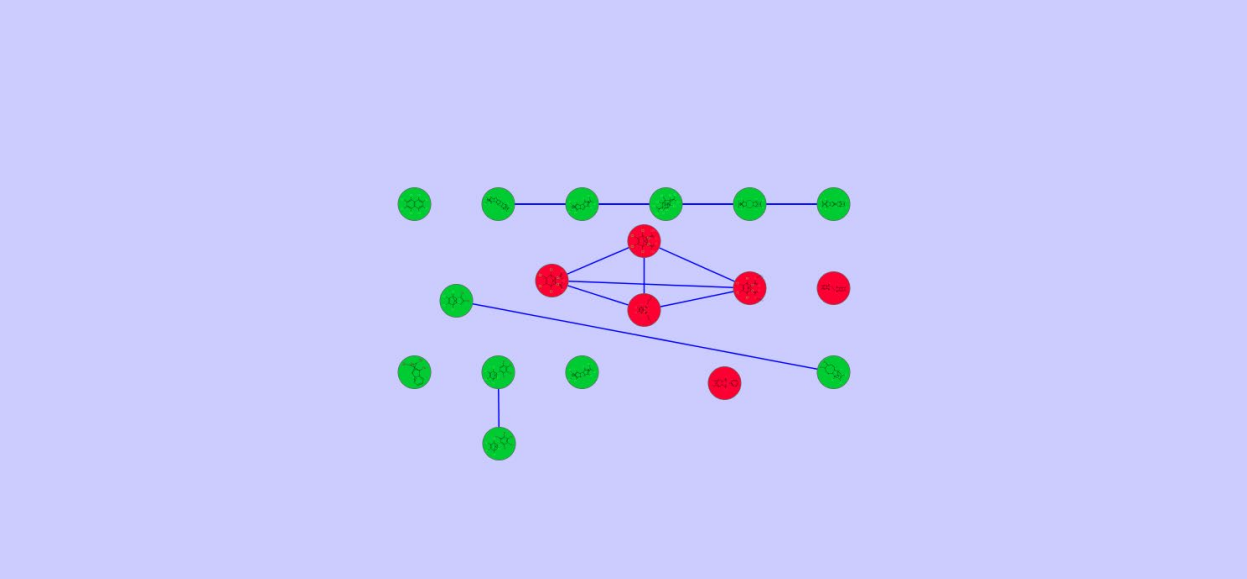

Figure 4B – All chemicals in the polyhalogenated carbocycles subclass; linked chemicals have a Tanimoto of 0.85; red indicates high GI availability.

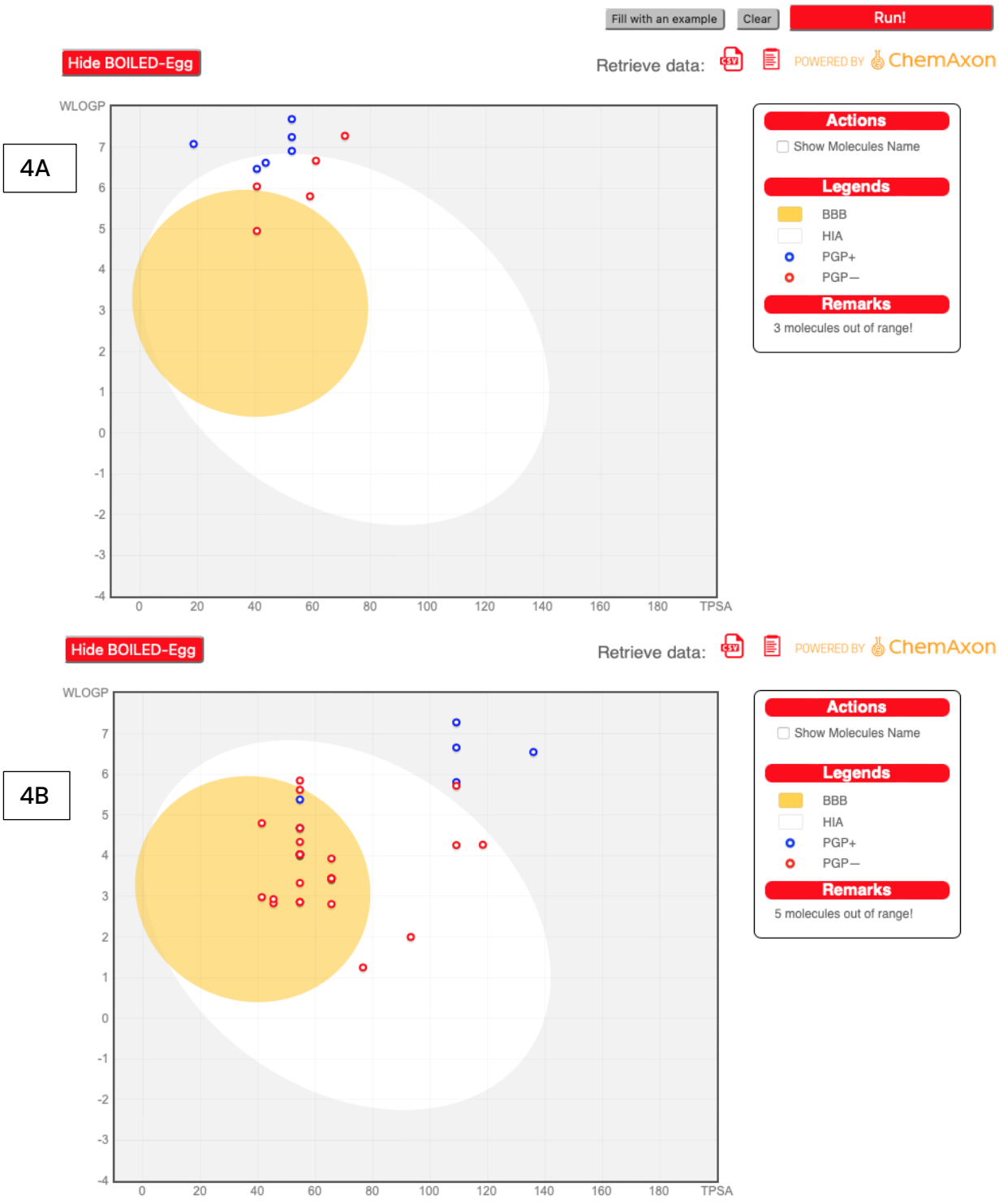

Figure 5 - SWISS ADME Prediction for (A) the polyhalogenated phenol aliphatic ether and (B) polyhalogenated organophosphate subclasses. Chemicals within the white space are predicted to have high intestinal absorption, within the yellow space are predicted to cross the blood brain barrier. PGP predictions (p-glycoprotein transporters) are noted in the graphic, but are not expected to be useful for this class.

Results from Task 1B

Our analysis of the OFRs gave us the following main takeaways:

- The best analog for an OFR may be another OFR from a different subclass.
- Data availability, bioavailability, metabolic pathway, and data gap analysis can likely all be used to trim the list of potential analogs in subsequent tasks.
- Analysis of physicochemical properties showed that the substances tend to be in a narrow range, which meant that distinct clusters could not be defined above and beyond the chemical similarity metric used above; however, further work may elucidate useful groupings.
- Analysis of bioavailability of the OFRs indicated that there is a wide range of bioavailability and therefore further evaluation of analog quality by bioavailability will depend on data from Task 2 on metabolic processes.

4. Characterization of Data Availability from Toxicity Databases for Screened Analogs (Task 1C)

Before developing a working set of analog chemicals for preliminary read-across analyses, we first determined data availability for the analogs. Data availability from toxicity databases was characterized for the trimmed and deduplicated analogs from Task 1A using evidence maps as developed for the literature survey compiled through previous CPSC task orders (#61320621F1001, 61320621F1002, and 61320621F1003 [Tasks #12 through 14] under contract CPSC-D-17-0001). Detailed documentation of the approach can be found in the Literature Survey Guide[1] .

Databases

We used the same database sources as those used for the literature survey. The one difference is that we focused on measured data instead of modeled data - i.e., the Danish QSAR Database/Models used in the literature survey were excluded here. Since the Multimedia Monitoring Database (MMDB) is focused on exposure, MMDB is also excluded for analogs. The final databases from which data were downloaded separately for each subclass used were:

- Environmental Protection Agency's (EPA) CompTox Chemicals Dashboard Batch Download, including data from ToxVal,
- National Toxicology Program's (NTP) Integrated Chemical Environment (ICE)
- Organisation for Economic Co-operation and Development's (OECD) Quantitative Structure-Activity Relationship (QSAR) Toolbox
- PubChem BioAssays
- Comparative Toxicogenomics Database (CTD)

[1] UC Risk Science Center (2022). Literature Survey Guide.

- European Chemicals Agency (ECHA) Dossier study results obtained via International Uniform Chemical Information Database (IUCLID) software
    - ECHA Dossier study results were only searched for chemicals that had data from the other databases in this list

In addition, the databases from which a single download was used to extract data for all subclasses were:

- Table S4 of the Supplementary Material from Vegosen and Martin (2020)[2], which replaces the "backdoor" CompTox downloads that were used in the literature survey.
- National Institute for Occupational Safety and Health (NIOSH) Pocket Guide
- Endpoint-specific datasets from ICE, available at: https://ice.ntp.niehs.nih.gov/DATASETDESCRIPTION
    - Acute Oral Toxicity
    - Cancer
    - Developmental and Reproductive Toxicity (DART)
    - Eye Irritation
    - Sensitization
    - Skin Irritation-Corrosion
    - ADME Parameter

One additional difference from the literature survey is that the final evidence maps do not include columns for data on Human Health Risk Assessment. Such data is not found in the toxicity databases, so is not included in the current evidence maps.

Chemical Identifiers in Database Downloads and Code

For both the (i) database downloads of toxicity data and (ii) extraction of downloaded data and assembly into evidence maps, the analogs in the present work have one key difference from the literature survey work. In the literature survey, all chemicals were identified with a CAS RN. In the present work, many analogs identified by PubChem and Chemquest do not have an associated CAS RN. Thus, (i) the process for downloading and (ii) the code for assembling evidence maps was updated to accommodate additional identifiers. Based on the identifiers that can be used to search the selected database sources, the PubChem CID, SMILES, and InChiKey were added as alternate identifiers. The updated evidence map code is provided with this set of deliverables.

Results from Task 1C

We created evidence maps for all 14 subclasses that describe the availability of toxicity data for the trimmed list of analogs. These files also include tabs listing all of the analogs for which no toxicity data were available. Our current overall lists of analogs without data do not capture these analogs as having no data. These lists of analogs with no data in the evidence map files

[2] Vegosen and Martin. (2020) An automated framework for compiling and integrating chemical hazard data. *Clean Technologies and Environmental Policy* 22, 441-458.

include both analogs identified as part of this task, as well as analogs from the literature survey. Note that there are a large number of analogs from the literature survey that only had data from Danish QSAR and nowhere else. Thus, the criteria for inclusion of an analog in the list is slightly different between literature survey analogs and analogs newly identified as part of this task. These evidence maps will be used to further trim the list of analogs developed in Task 1A. Table 2 presents the number of analogs after trimming for data availability.

Table 2. Total Number after (i) Trimming and Deduplication from Task 1A and (ii) Trimming Based on Data Availability

No.	Subclass (all Polyhalogenated)	Abbreviation	Total after Trimming and Deduplication (Task 1A)	Total after Trimming for Data Availability (Task 1C)
1	Organophosphates	PHOP	1,519	144
2	Bisphenol aliphatics and functionalized	PHBAF	7,797	196
3	Diphenyl ethers	PHDE	2,401	122
4	Benzene alicycles	PHBA	56	5
5	Benzene aliphatics and functionalized	PHBzAF	5,751	266
6	Phenol aliphatic ethers	PHPhAE	5,656	168
7	Alicycles	PHA	662	38
8	Phenol derivatives	PHPhD	1,016	33
9	Aliphatic carboxylates	PHACbx	347	20
10	Benzenes	PHB	6,590	150
11	Phthalates-benzoates-imides	PHPBI	3,916	214
12	Carbocycles	PHC	1,433	116
13	Triazines	PHT	111	4
14	Aliphatic chains	PHACh	1,185	51

5. Next Steps

Next steps will include to continue to refine the lists of potential analogs by incorporating data availability, which is the critical criterion, and then by matching predicted ADME properties, and finally by identified metabolic pathways or identified reactive functional groups that can connect toxicological activity to structural motifs. In addition, based on our clustering by chemical similarity, several subclasses lend themselves to read-across and will not require a great deal of data gap filling beyond the existing OFRs, while others will be more dependent on identified analogs. This will allow us to further trim the lists to focus on those compounds that have data, and to develop methods to identify and fill data gaps.