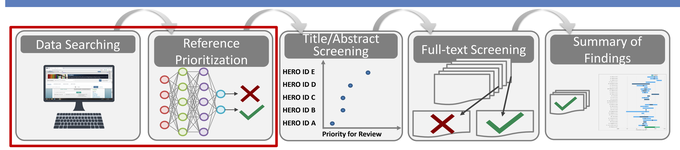
Evidence Mapping for Engineering & Exposure: Literature Search, Prioritization and Pre-Screening Strategy

Ariel Hou, Yadi Lopez, Nerija Orentas, Chantel Nicolas, Katherine Phillips, Yvette Selby-Mohamadu
U.S. EPA, OCSPP/OPPT/RAD, Washington, D.C.

www.epa.gov

Ariel Hou | Hou.ariel@epa.gov | 202-564-5591

General Workflow for Engineering & Exposure Evidence Mapping



The **Engineering & Exposure Evidence Mapping Workflow** starts with a comprehensive search of peer-reviewed literature databases using chemical names (including synonyms) to identify the literature pool for systematic review. The search results are deduplicated using EPA's HERO database. After deduplication, the literature pool is prioritized using SWIFT Review to narrow down to a smaller set of references likely to be relevant for Exposure before they undergo Title/Abstract Screening in SWIFT Active Screener.

Scope of Engineering & Exposure under TSCA Systematic Review:

Engineering
- Occupational exposure
- Environmental releases

Exposure
- Environmental exposure
- General population exposure
- Consumer exposure

Databases searched for Next 20 High Priority Substances:

- Agricola
- Dissertation abstracts
- PubMed (National Library of Medicine)
- Science Direct
- TOXNET
- ECOTOX UNIFY
- Web of Science (Thomson Reuters)

Step 1: Collecting Positive and Negative Seed References for Reference Prioritization

SWIFT Review is a literature review classification software used by EPA for reference prioritization. The software requires both positive and negative seeds to "rank" the literature pool. References whose titles and abstracts most closely resemble the positive seed articles are ranked higher in the prioritization process.

- **Positive Seeds** are the title and abstract of references known to contain relevant information for the discipline of interest
- **Negative Seeds** are the title and abstract of references known to NOT contain relevant information

To identify **Positive Seeds**, EPA used the exposure literature pool for the first 10 TSCA Risk Evaluations. The positive seed references were those that supported technical aspects of the exposure assessment for the 1-bromopropane, cyclic aliphatic bromide cluster (HBCD), methylene chloride, N-methylpyrrolidone (NMP), perchloroethylene, trichloroethylene, and asbestos draft TSCA Risk Evaluations.

Table 1. Number of Positive Seed References from the First Ten TSCA Risk Evaluation Used for Reference Prioritization

Chemical	Number of Positive Seeds	
	Engineering	Exposure
1-Bromopropane	7	9
Asbestos	8	7
Cyclic aliphatic bromide cluster	6	378
Methylene chloride	9	8
n-methylpyrrolidone	5	0
Trichloroethylene	2	6
Perchloroethylene	6	59
Other (covers multiple chemicals)	7	16
Total	**50**	**483**

Note:
Engineering coverings Occupational Exposure and Environmental Release
Exposure covers Environmental, General Population, and Consumer Exposure

Table 2. Number of Positive Seeds by Data Element Used for Reference Prioritization for the Engineering and Exposure Disciplines

Engineering Data Type	Number of Positive Seeds	
	Engineering	Exposure
General Facility Estimate	1	n.a.
Occupational Exposure	40	n.a.
Environmental Release	4	n.a.
Multiple	5	27
Consumer	n.a.	75
Dietary	n.a.	24
Environmental Exposure	n.a.	311
Human Biomonitoring	n.a.	46
Total	50	**483**

n.a. – Not applicable

Negative Seeds were selected using the following method for Engineering and Exposure:

Engineering –
- 50 negative seeds for each set of references to be prioritized
- Same number as positive seeds for most optimal prioritization
- Manually selected based on review of title/abstract determined to be least relevant to the data element of interest

Exposure –
- 473 negative seeds, selected from six compound of the next 20 compounds (one from each compound group)
- Roughly the same total number of negative seeds as positive seeds
- Manually selected based on review of title/abstract to be irrelevant to exposure

Step 2: Assessing the Performance of Reference Prioritization Method

To assess performance of the Reference Prioritization Method, validation test runs and/or analyses were performed to ensure that the positive seeds are capable of capturing relevant information and the negative seeds are capable of identifying references with no relevant information.

For ***Engineering*** (occupational exposure & environmental release), a total of 5 validation test runs were performed using the selected positive seeds to score a known set of literature references in SWIFT Review. Specifically:

- Positive seeds were used to numerically score references tagged for the draft 1,4-dioxane, HBCD, 1-BP, NMP, and methylene chloride Risk Evaluations in SWIFT Review
- Scores were reviewed to make sure that the Engineering integrated references (i.e., those that supported technical engineering aspects of the draft Risk Evaluation) received a higher score relative to other references that were not used or were not integrated
- Generally, the validation test runs show that all integrated references from the known datasets scored at the 80th percentile or higher.
- From these results, EPA determined the 80th percentile score as the "*cut-off score*". Prioritized references that score above this cut-off will move forward to Title/Abstract Screening

For ***Exposure*** (environmental, general population, and consumer exposure), 5-fold cross validation was performed. The positive and negative seeds were split into five folds; SWIFT-Review scoring was carried out 5 times, each time the scoring is trained on 4 of the 5 groups of seeds and the held out group is scored:

- Positive and negative seeds were reviewed to ensure they were properly scored (positive seeds had high scores while negative seeds had low scores). The lowest positive seed score was 0.7; the highest negative seed score was 0.37
- This cross-fold validation exercise shows that SWIFT-Review can discern between the selected positive and negative exposure seeds
- The "*cut-off score*" for deciding if a reference should be carried forward to SWIFT-Active Screener was determined by subtracting two standard deviations of the distribution of positive seed scores from the minimum positive seed score

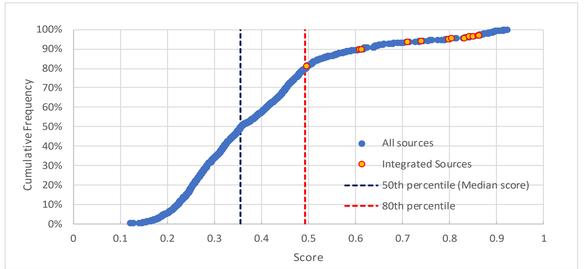


Figure 1. Cumulative Frequency of SWIFT Review Scores from the 1-BP Validation Test Run (Reference Dataset from the draft 1-BP TSCA Risk Evaluation)

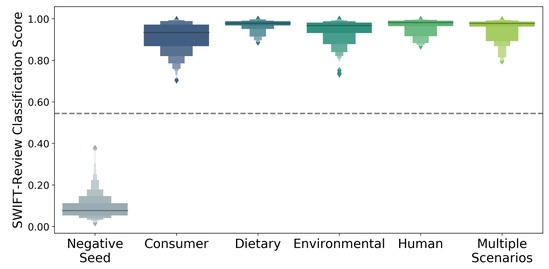


Figure 2. Distributions of SWIFT-Review scores for positive seeds split by different exposure scenarios and the scores for the negative exposure seeds. The dotted grey line shows the cutoff that can be applied to determine if a scored reference would be sent on to SWIFT-Active Screener.

Step 3: Screening References in SWIFT-Active Screener (Active Machine Learning)

After Step 2, the prioritized references undergo Title and Abstract Screening in SWIFT-Active Screener. **SWIFT-Active Screener** is a web-based, collaborative systematic review software application that EPA adopted for the TSCA Systematic Review for the Next 20 High Priority Substances.

The software uses an active machine learning algorithm where, as screeners include or exclude references, it periodically computes which and how many of the remaining unscreened references are most likely to be relevant. Using this software allows EPA to manually screen only a portion of the prioritized references, focusing its resources on those that are most likely to be relevant to TSCA Risk Evaluations.

Each reference is reviewed by two screeners against a chemical-agnostic **R**eceptor, **E**xposure, **S**etting/Scenario, and **O**utcome (**RESO**) statement, and conflicts are resolved by a third, independent screener.