A public database supporting evidence-based exposomics

Risa R. Sayre[1,2], John F. Wambaugh[1], Katherine Phillips[3], Antony J. Williams[1], Christopher M. Grulke[1]

1) U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, RTP NC 2) Oak Ridge Institute of Science and Education (ORISE) Research Participant 3) U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Computational Exposure Division, RTP NC

Risa R. Sayre | ORCID 0000-0002-6173-8020 | sayre.risa@epa.gov

www.epa.gov

1) Background

To support **identification** of likely **sources** of chemicals found in biological media through non-targeted/suspect screening analysis (SSA/NTA), this work-in-progress annotates chemicals with likely origin categories and adds **empirically-validated substance relationships** between chemicals and their *in vivo* **transformation products** to the CompTox Chemicals Dashboard[1].

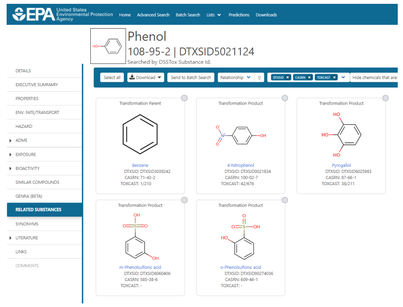
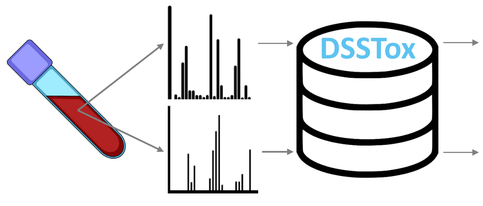


Fig 1. Screenshot of the **Related Substances** tab for phenol in the CompTox Chemicals Dashboard. One parent and four transformation products have been registered.

2) Method: Categorizing chemical origin

We identify five categories of chemical origin (based on Rappaport *et al* 2014[2]) of small molecules found in human blood biomonitoring samples: 1) **endogenous** metabolome, 2a) exogenous **nutrients**, 2b) markers of exposure to exogenous nutrients, 3a) **xenobiotics** (pharmaceuticals, pesticides, and others), and 3b) markers of exposure to xenobiotics.



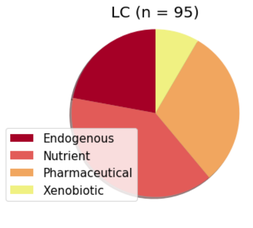
Endogenous: will be observed in all human sera, regardless of exposures

Transformation product from xenobiotic: where did this come from?

To group chemicals into these categories, a one vs. one linear support vector classifier was trained on the URLs of the top ten Google results for chemical names from manually curated and Dashboard-registered lists for categories 1 (from Rappaport *et al* 2014 supplement), 2a (from FDA Substances Added to Food[3]), and 3a: pesticides[4], pharmaceutical active ingredients, and other (TSCA[5], with overlapping names from other categories removed). The training set (n = 2640) was not restricted to compounds observed in blood. The overall F1 score on the validation set (n = 1320) of the model was 0.80.

3) Case: Categories observed in a pooled blood sample

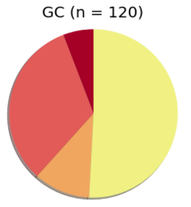
Compounds from pooled human serum samples were tentatively identified (in at least 2 out of 3 replicates) in GCxGC-MS and LC-QTOF SSA/NTA workflows (complete methods to be described in a future publication with Lesa Aylward of Summit Toxicology) designed to filter out endogenous compounds[6]. 22% of the compounds identified in the LC workflow were not registered in DSSTox (EPA's Distributed Structure-Searchable Toxicity Database[7]), most of which were endogenous.

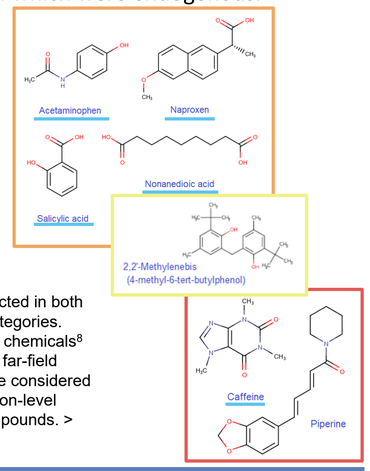


< Proportions of categories observed in LC and GC.

GC workflow: Katherine Phillips, Alice Yau (Southwest Labs)

LC workflow: Alex Chao, Jon Sobus, Barbara Wetmore



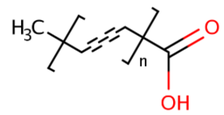


All the compounds detected in both LC and GC and their categories. High production volume chemicals[8] are underlined; near- or far-field exposure factors may be considered when modeling population-level exposures of these compounds. >

4) Method: Adding chemicals from EPA-relevant exposomics categories to DSSTox library

Pre-filtering categories 1 & 2

To avoid registering non-xenobiotic compounds, we created **chemical structural classes** to pre-filter chemicals from the identification workflow.



Markush query for "fatty acids"

Addition of category 3b compounds

After finding a high number of false positives (>99%) in a PubMed search for "metabolite of [xenobiotic name]", we used manually classified abstracts to build a natural language processing model (F1 = 0.98) to identify abstracts containing substrate/product pairs, or **substance relationships**. 74% of these transformation products were previously unregistered in DSSTox. To increase signal without adding noise, we registered only transformations **observed** at plausible exposure levels (and not rapidly transforming intermediates), which are linked to detection method and other metadata.

5) Method: Supporting NTA identification

Existing capability

Advanced mass- and formulae-based searches in the Dashboard, including consideration of adducts[9]. Ranking of candidates utilizes predicted fragmentation patterns and metadata

\+ ***New metadata from this project***

- Structures grouped by multiple **chemical lists** of observed compounds in environmental and biological media support NTA
- **Substance relationship mappings** allow metadata aggregation (such as data source counts) from known transformation parents to their children, possibly improving proper identification of children (*related families*)

6) Discussion

Over 10,000 mappings of xenobiotic transformation relationships are being added to DSSTox, many of which are not currently registered in any metabolomics database. Developing methods to improve identification of these substances measured in human blood and their source categories supports active Agency research projects (e.g. for PFAS chemicals).

Registration of xenobiotics and observed transformation products based on dose levels demonstrated to yield a detectable (by a known method) amount of product in a particular species and medium in a chemical library

- allows **development of exposure estimates**
- identifies **candidate substances and pathways** to inform future **high-throughput assay research** to identify mechanisms

7) References

1) Williams AJ, *et al*. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform*. 2017 Nov 28;9(1):61. ***2)*** Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The Blood Exposome and Its Role in Discovering Causes of Disease. *Environ Health Perspect.* 2014 Aug; 122(8). ***3)*** U.S. Food & Drug Administration. Substances Added to Food (formerly EAFUS). www.accessdata.fda.gov/scripts/fdcc/?set=FoodSubstances ***4)*** U.S. EPA Office of Pesticide Programs. Pesticide Chemical Search. iaspub.epa.gov/apex/pesticides/f?p=chemicalsearch:1 ***5)*** U.S. EPA TSCA Chemical Substance Inventory. www.epa.gov/tsca-inventory/how-access-tsca-inventory ***6)*** Sobus JR, *et al.* Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J Expo Sci Environ Epidemiol*. 2017 Dec 29. ***7)*** Richard AM, Williams CR. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res*. 2002 Jan 29;499(1):27-52. ***8)*** U.S. EPA High Production Volume Information System. http://iaspub.epa.gov/oppthpv/hpv_hc_characterization.get_report?doctype=2 ***9)*** McEachran AD, Sobus JR, Williams AJ. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal Bioanal Chem*. 2017 Mar;409(7):1729-1735.

This project was supported in part by an appointment to the Internship/Research Participation Program at the National Center for Computational Toxicology, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.

This poster does not necessarily reflect U.S. EPA policy.