High-throughput literature mining to support read-across predictions of toxicity

Nancy C. Baker, PhD
Leidos, contractor to the US EPA

September 29, 2016

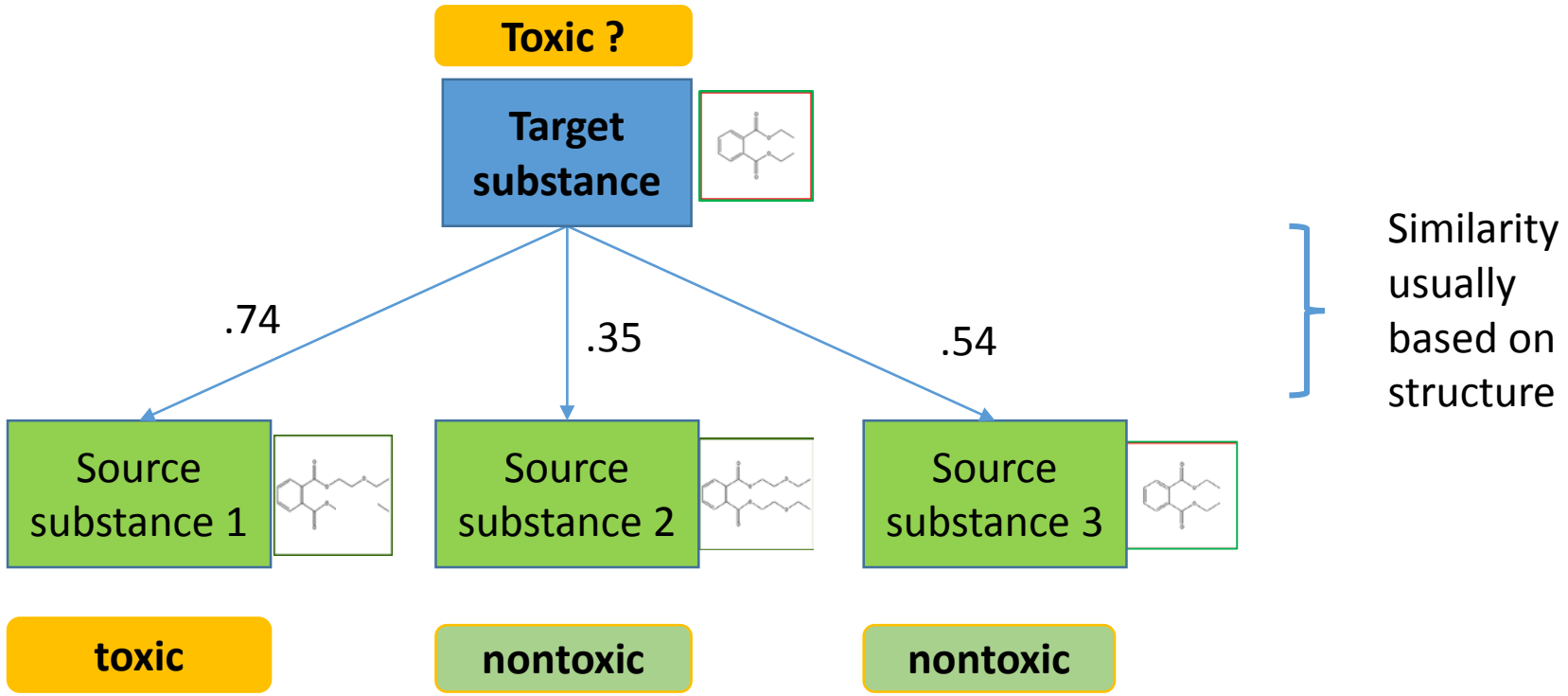
ASCCT

DISCLAIMER: The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Outline

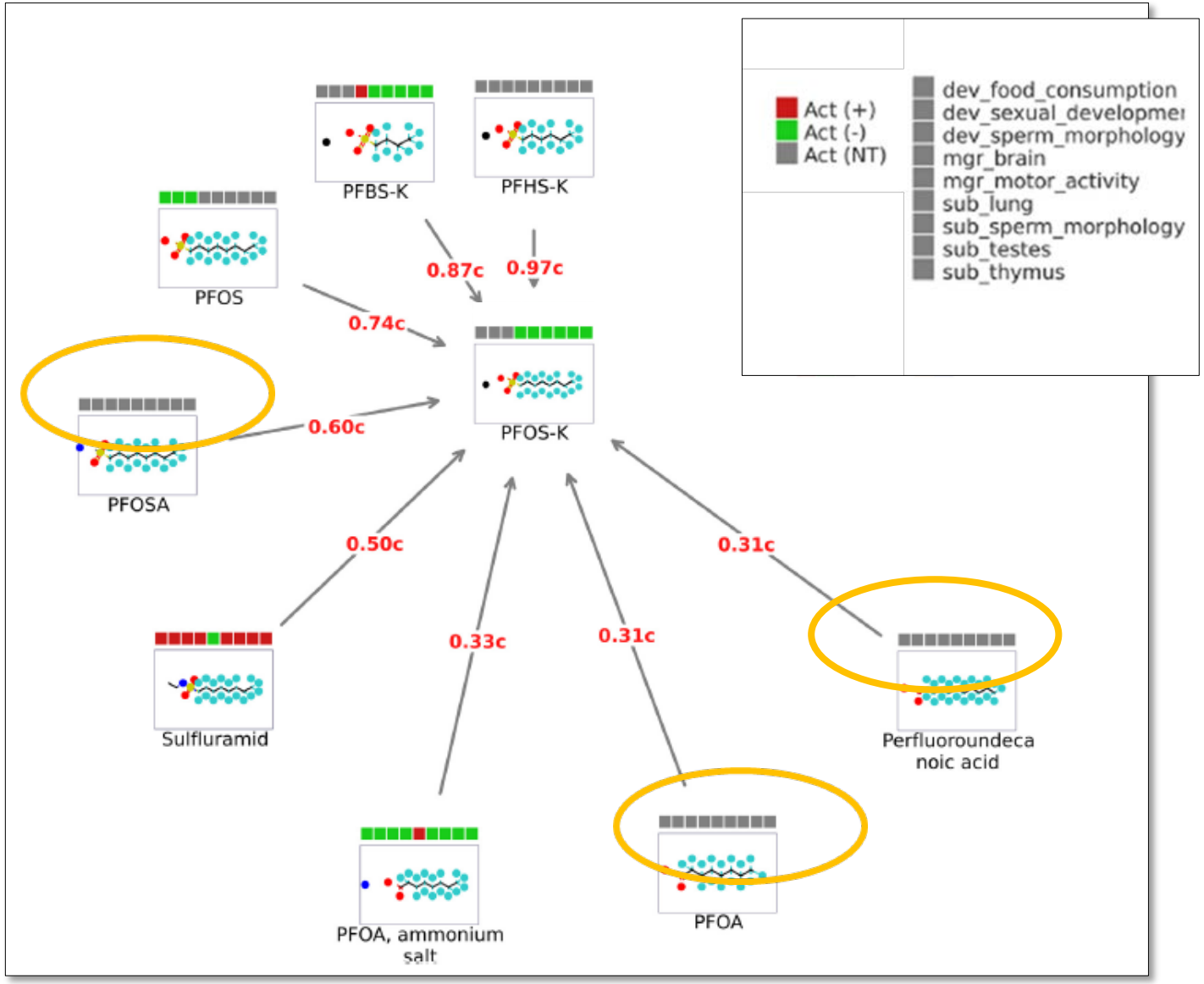
- Why literature mining?
- Methods to gather and process information from the biomedical literature
- Examples of how this solution can be applied to read-across

Read-across: technique for filling data gaps



Example: Filling the information gap with ToxRefDB animal assay data

Problem: ToxRefDB data is not available for all chemicals or for all endpoints.



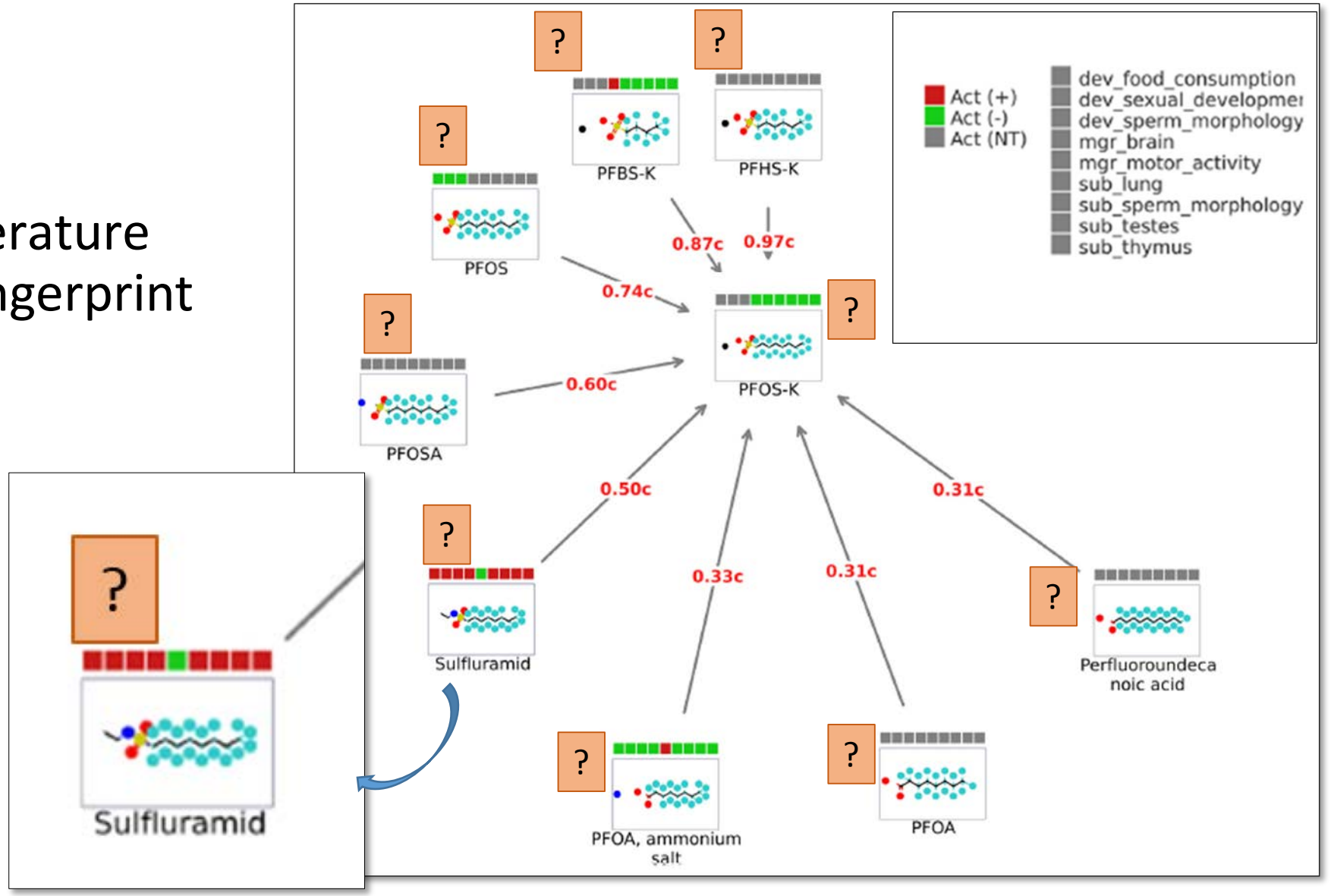
Modified from Shah, I. et al. Regulatory Toxicology and Pharmacology, Volume 79, 2016,12-24

Literature can inform read-across

- Why the literature?
  - Large and growing source
    - > 24 million articles; > 12 million about chemicals
    - Encompasses all sorts of toxicity
- Challenges
  - Literature is unstructured – not data
  - Limitations of literature mining; e.g., publication bias and granularity
  - Large and growing source – challenges for human cognition of big data
- Goal: gather and condense literature information into a signature that can be visualized for read-across.

Goal

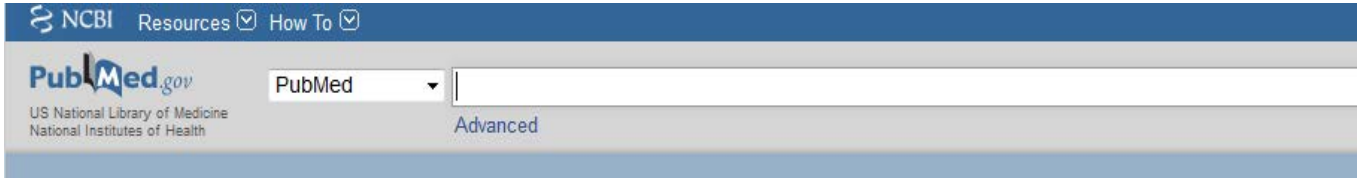
- Construct a literature signature or fingerprint or descriptors
- Visualize



Methodology for literature mining to inform read-across

1. Gathering and extracting the literature information
2. Organizing by toxicity type
3. Condensing and strengthening signal → signature
4. Visualizing

1. Gathering and extracting



Format: Abstract

Send to

Biochem Pharmacol. 1993 Oct 19;46(8):1385-91.

Hexachlorobenzene-induced hypothyroidism. Involvement of different mechanisms by parent compound and metabolite.

van Raaij JA[1], Frijters CM, van den Berg KJ.

Author information

Abstract

Rats received repeated oral treatment with different doses of hexachlorobenzene (HCB) (0-3.5 mmol/kg) for 2 or 4 weeks. Measurements of thyroid hormone status after 2 weeks showed a dose-dependent decrease of total thyroxine (TT4) levels, decreased free thyroxine (FT4) levels and little change of total triiodothyronine (TT3) levels. The effects on thyroid hormone status were more pronounced after 4 weeks and also included increased thyroid stimulating hormone (TSH) levels. These conditions suggest that HCB had induced hypothyroidism in these animals. Indications for occupation of thyroid hormone binding proteins were found in serum of exposed animals. The major metabolite pentachlorophenol (PCP) also caused, by competitive interactions with thyroid hormone binding proteins in serum, a rapid and dose-dependent decrease of TT4 and FT4 levels, but not of TT3 levels in serum. The decrease of serum TT4 levels by repeated dosing with 3.5 mmol HCB/kg for 4 weeks could be attributed to competitive interactions of PCP with hormone serum binding proteins and to increased metabolism induced by HCB to an equal degree. At lower dose levels or with shorter dosing periods, increased metabolism of T4 is the main cause of decreased TT4 serum levels. This is the first indication that a similar effect is caused simultaneously by the parent compound and its metabolite through different and independent mechanisms.

PMID: 8240387

[PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances

Publication Types, MeSH Terms, Substances

Publication Types

Research Support, Non-U.S. Gov't

MeSH Terms

- Animals
- Binding, Competitive
- Blood Proteins/metabolism
- Body Temperature/drug effects
- Dose-Response Relationship, Drug
- Hexachlorobenzene/blood
- Hexachlorobenzene/metabolism
- Hexachlorobenzene/toxicity*
- Hypothyroidism/chemically induced*
- Hypothyroidism/metabolism
- Male
- Pentachlorophenol/administration & dosage
- Pentachlorophenol/blood
- Rats
- Rats, Wistar
- Thyroid Hormones/metabolism
- Thyroxine/blood

National Library of Medicine Indexers

Indexing terms → data

PubMed ID	MeSH heading	Qualifier / subheading	Major topic?
8240387	Hexachlorobenzene	Toxicity	Y

PubMed ID	MeSH heading	Qualifier / subheading	Major topic?	Score
8240387	Hypothyroidism	Chemically induced	Y	2
8240387	Body Temperature	Drug effects	N	2
8240387	Thyroid Hormones	Metabolism	N	1
8240387	Thyroxine	Blood	N	1

Score reflects confidence.

High-throughput text-mining: a few readouts per article, but it adds up …

Hexachlorobenzene – 1485 articles

348 biological processes

Biological processes	Article Count
Organ Size	73
Body Weight	62
Enzyme Induction	36
Reproduction	17
Immunity	11
Birth Weight	6
Oxygen Consumption	5
Phagocytosis	5
Overweight	5
Motor Activity	4
Weight Gain	4
Cell Proliferation	4
Oxidative Stress	4
Oxidative Phosphorylation	4
Phosphorylation	4
Gluconeogenesis	4
Fertility	4
Apoptosis	4
Child Development	3
Obesity	3
Homeostasis	3
Lipid Peroxidation	3
Gene Expression	3

269 Proteins / genes

Protein / gene	Article Count
Cytochrome P-450 Enzyme System	81
Uroporphyrinogen Decarboxylase	54
Carboxy-Lyases	39
Cytochrome P-450 CYP1A1	24
5-Aminolevulinate Synthetase	21
porphyrinogen carboxy-lyase	18
Glutathione	17
Thyroxine	16
Mixed Function Oxygenases	15
Aryl Hydrocarbon Hydroxylases	15
Receptors, Aryl Hydrocarbon	15
Glutathione Transferase	12
Oxygenases	11
Aminolevulinic Acid	11
Aminopyrine N-Demethylase	11
Triiodothyronine	11
Immunoglobulin M	11
Ferrochelatase	9
Immunoglobulin G	9
Receptors, Estrogen	8
Aniline Hydroxylase	8
7-Alkoxycoumarin O-Dealkylase	8
gamma-Glutamyltransferase	8
Alanine Transaminase	6

180 Diseases / conditions

Diseases	Article Count
Porphyrias	184
Body Weight	87
Drug-Induced Liver Injury	36
Prenatal Exposure Delayed Effects	30
Disease Models, Animal	27
	26
nental	22
	21
	16
	14
	12
	11
l	10
	8
	7
ar	6
	6
	5
	5
	5
	5
	5
	5
ıced	5
	5
	5
	4
	4
	3
ons	3

185 Anatomical terms

Anatomy Terms	Article Count
Liver	286
Adipose Tissue	124
Milk, Human	74
Microsomes, Liver	67
Feces	45
Kidney	39
Milk	27
Thyroid Gland	23
Skin	23
Brain	22
Lung	21
Fetal Blood	20
Muscles	19
Spleen	19
Mitochondria, Liver	17
Fetus	14
Bile	14
Ovary	12
Ovum	11
Chick Embryo	11
Placenta	11
T-Lymphocytes	11
Macrophages	10
Erythrocytes	10
Thymus Gland	9
Intestines	9
Lymph Nodes	8
Myocardium	8

Gathering: how big is the data?

- 24 million articles in PubMed
- 12 million articles have chemical annotations
- 200 million MeSH annotations
- Growth rate: 1 million / month
- When looking only at annotations of most interest …
  - 65 million chemical annotations
  - 61 million disease / effect annotations (diseases, anatomy)

2. Defining toxicity type and organizing by it

Toxicity Type
Thyroid
ReproTox
GeneTox
DevTox
…

Specify detail

Toxicity Type	MeSH Terms	Categories
Thyroid	Hypothyroidism	Clinical conditions
	Hyperthyroidism	Clinical conditions
	Hashimoto Disease	Clinical conditions
	Thyroid gland	Thyroid gland
	Thyroid nodule	Thyroid gland
	Receptor, Thyroid Hormone	Receptors
	Receptor, Thyrotropin	Receptors
	Thyroxine	Hormones
	Triiodothyronine	Hormones

Subject matter experts advise on the terms and categories for each Toxicity Type.

2. Organizing by toxicity type

Chemical	PubMed ID	MeSH heading	Score	ToxType	Category	Category Score
Hexachlorobenzene	8240387	Hypothyroidism	2	Thyroid	Clinical Conditions	2
	8240387	Body Temperature	2	Thyroid	Body Temp Regulation	2
	8240387	Thyroid Hormones	1	Thyroid	Hormones	2
	8240387	Thyroxine	1	Thyroid	Hormones	

Next step: 3. Condensing into a signature.

Having a score that can be summarized is key.

3. Condensing into signature

Each numeric value is the total score for that chemical for the corresponding toxicity category / type.

Overview

		Scores (annotation count * arbitrary weights)																																		
		Androgen			DevTox			GeneTox				Obesity				ReproTox						Thyroid															
Chemical	cas	Agonism	Antagonism	Other receptor effect	Abnormality	Embryonic Structures	Morphogenesis	Mutagen	DNA Damage/Repair	Genetic Structures	Processes	Adipogenesis	Adipose Tissue	Clinical Conditions	Proteins and genes	Abnormality	Cancer	FemaleRepro	Infertility	MaleRepro	Processes	Binding Proteins	Hormones	Receptors	Synthesis	Hepatic Catabolism	TR Contolled Genes	Transporters	Body Temp Reg	Clinical Conditions	Cognition_IQ	Frog	Thyroid Gland
Benzene	71-43-2	0	0	0	1	4	2	78	186	27	1	0	0	8	0	0	0	0	0	0	8	0	1	0	5	4	0	0	0	0	0	0	0	0	0	0
3-dinitrobenzene	99-65-0	0	0	0	0	0	0	3	0	3	0	0	0	5	0	0	0	2	8	66	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hexachlorobenzene	118-74-1	0	0	0	0	1	0	6	7	0	0	0	2	11	0	0	8	8	0	0	22	1	12	1	1	4	0	0	1	4	3	0	4	0	0	0
nitrobenzene	98-95-3	0	0	0	0	1	0	18	21	4	0	0	0	5	0	0	0	0	4	20	15	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
4-dichlorobenzene	106-46-7	1	0	1	9	2	2	15	12	6	1	0	2	9	0	0	0	5	0	1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p-Aminoazobenzene	60-09-3	1	0	1	1	0	0	42	12	19	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Dinitrochlorobenzene	97-00-7	0	0	0	1	1	0	27	0	10	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
p-Dimethylaminoazobenzene	60-11-7	0	0	0	0	0	0	15	3	12	0	0	0	1	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0
2-dichlorobenzene	95-50-1	0	0	0	3	2	1	3	6	4	1	0	0	5	0	0	0	0	0	0	6	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
1,2-diaminobenzene	615-28-1	0	0	0	0	0	0	15	12	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
azobenzene	103-33-3	0	0	0	0	1	0	12	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
dodecylbenzenesulfonic acid	27177-77-1	0	0	0	3	2	1	6	6	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3-methoxy-4-aminoazobenzene	3544-23-8	1	0	1	0	0	0	12	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,2,4,5-tetrachlorobenzene	95-94-3	0	0	0	1	2	0	0	0	0	0	0	1	2	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4-chloro-1,2-diaminobenzene	95-83-0	0	0	0	0	0	0	12	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Double-click on signature scores goes back to detail.

Hexachlorobenzene	Thyroid_Binding Proteins	Prealbumin	1	7801323	1994	Reduction of thyroxine uptake into cerebrospinal fluid and rat b
Hexachlorobenzene	Thyroid_BodyTempReg	Body Temperature	1	8240387	1993	Hexachlorobenzene-induced hypothyroidism. Involvement of d
Hexachlorobenzene	Thyroid_Clinical Conditions	Hypothyroidism	2	1755017	1991	Hexachlorobenzene and its metabolites pentachlorophenol and
Hexachlorobenzene	Thyroid_Clinical Conditions	Hypothyroidism	2	8240387	1993	Hexachlorobenzene-induced hypothyroidism. Involvement of d
Hexachlorobenzene	Thyroid_Cognition_IQ	Cognition	3	23085522	2013	Evaluating the neurotoxic effects of lactational exposure to pers
Hexachlorobenzene	Thyroid_Hepatic Catabolism	Arylsulfotransferase	1	24365113	2014	Hexachlorobenzene and pentachlorobenzene accumulation, me
Hexachlorobenzene	Thyroid_Hepatic Catabolism	Glucuronosyltransferase	1	416059	1978	Enhancement of the UDP glucuronyltransferase activity and bili
Hexachlorobenzene	Thyroid_Hepatic Catabolism	Glucuronosyltransferase	1	8442763	1993	Increased glucuronidation of thyroid hormone in hexachlorober

Double-click brings the user here.

Methodology for literature mining to inform read-across

1. Gathering and extracting the literature information
2. Organizing by toxicity type
3. Condensing and strengthening signal → signature
4. Visualizing - use benzenes as example

Now we need visualization.

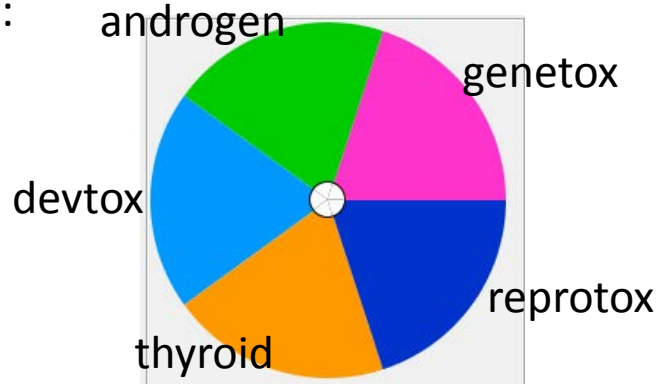
Let's start by selecting only benzenes.

Overview

		Scores (annotation count * arbitrary weights)																																		
		Androgen			DevTox			GeneTox				Obesity				ReproTox					Thyroid																
Chemical	cas	Agonism	Antagonism	Other receptor effect	Abnormality	Embryonic Structures	Morphogenesis	Mutagen	DNA Damage/Repair	Genetic Structures	Processes	Adipogenesis	Adipose Tissue	Clinical Conditions	Proteins and genes	Abnormality	Cancer	FemaleRepro	Infertility	MaleRepro	Processes	Binding Proteins	Hormones	Receptors	Synthesis	Hepatic Catabolism	TR Contolled Genes	Transporters	Body Temp Reg	Clinical Conditions	Cognition_IQ	Frog	Thyroid Gland
Benzene	71-43-2	0	0	0	1	4	2	78	186	27	1	0	0	8	0	0	0	0	0	0	8	0	1	0	5	4	0	0	0	0	0	0	0	0	0	0
3-dinitrobenzene	99-65-0	0	0	0	0	0	0	3	0	3	0	0	0	5	0	0	0	2	8	66	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hexachlorobenzene	118-74-1	0	0	0	0	1	0	6	7	0	0	0	2	11	0	0	8	8	0	0	22	1	12	1	1	4	0	0	1	4	3	0	4	0	0	0
nitrobenzene	98-95-3	0	0	0	0	1	0	18	21	4	0	0	0	5	0	0	0	0	4	20	15	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
4-dichlorobenzene	106-46-7	1	0	1	9	2	2	15	12	6	1	0	2	9	0	0	0	5	0	1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p-Aminoazobenzene	60-09-3	1	0	1	1	0	0	42	12	19	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Dinitrochlorobenzene	97-00-7	0	0	0	1	1	0	27	0	10	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
p-Dimethylaminoazobenzene	60-11-7	0	0	0	0	0	0	15	3	12	0	0	0	1	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0
2-dichlorobenzene	95-50-1	0	0	0	3	2	1	3	6	4	1	0	0	5	0	0	0	0	0	0	6	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
1,2-diaminobenzene	615-28-1	0	0	0	0	0	0	15	12	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
azobenzene	103-33-3	0	0	0	0	1	0	12	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
dodecylbenzenesulfonic acid	27177-77-1	0	0	0	3	2	1	6	6	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3-methoxy-4-aminoazobenzene	3544-23-8	1	0	1	0	0	0	12	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,2,4,5-tetrachlorobenzene	95-94-3	0	0	0	1	2	0	0	0	0	0	0	1	2	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4-chloro-1,2-diaminobenzene	95-83-0	0	0	0	0	0	0	12	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
trichlorobenzene	87-61-6	0	0	0	0	0	1	3	3	0	0	0	0	5	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Methyldimethylaminoazobenzene	55-80-1	0	0	0	0	0	0	6	0	2	0	0	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
chlorobenzene	108-90-7	1	0	1	0	2	0	0	3	0	1	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,4-bis(2-(3,5-dichloropyridyloxy))benzene	76150-91-9	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	4	0	1	0	0	0	0	2	0	0	0
pentachlorobenzene	608-93-5	0	0	0	0	0	1	0	0	0	0	0	0	3	0	0	0	1	0	0	2	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0
1,2,4-trichlorobenzene	120-82-1	0	0	0	0	1	1	0	3	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2-methoxy-4-aminoazobenzene	80830-39-3	0	0	0	0	0	0	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dinitrofluorobenzene	70-34-8	0	0	0	0	0	0	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N-hydroxy-4-aminoazobenzene		0	0	0	0	0	0	6	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,3-dichlorobenzene	541-73-1	0	0	0	0	2	0	0	0	0	1	0	0	2	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2-chloronitrobenzene	88-73-3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4-chloronitrobenzene	100-00-5	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N-methyl-4-aminoazobenzene		0	0	0	0	0	0	3	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1,2-diamino-4-nitrobenzene	99-56-9	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
sym-trinitrobenzene	99-35-4	0	0	0	0	1	1	3	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2,5-dichloronitrobenzene	89-61-2	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Visualization: new concept- LitToxPI

Key:



Example:

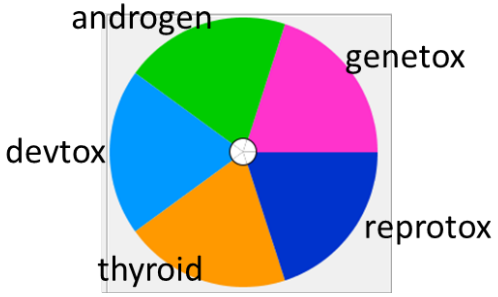
Hexachlorobenzene

Pie slice size reflects proportion.

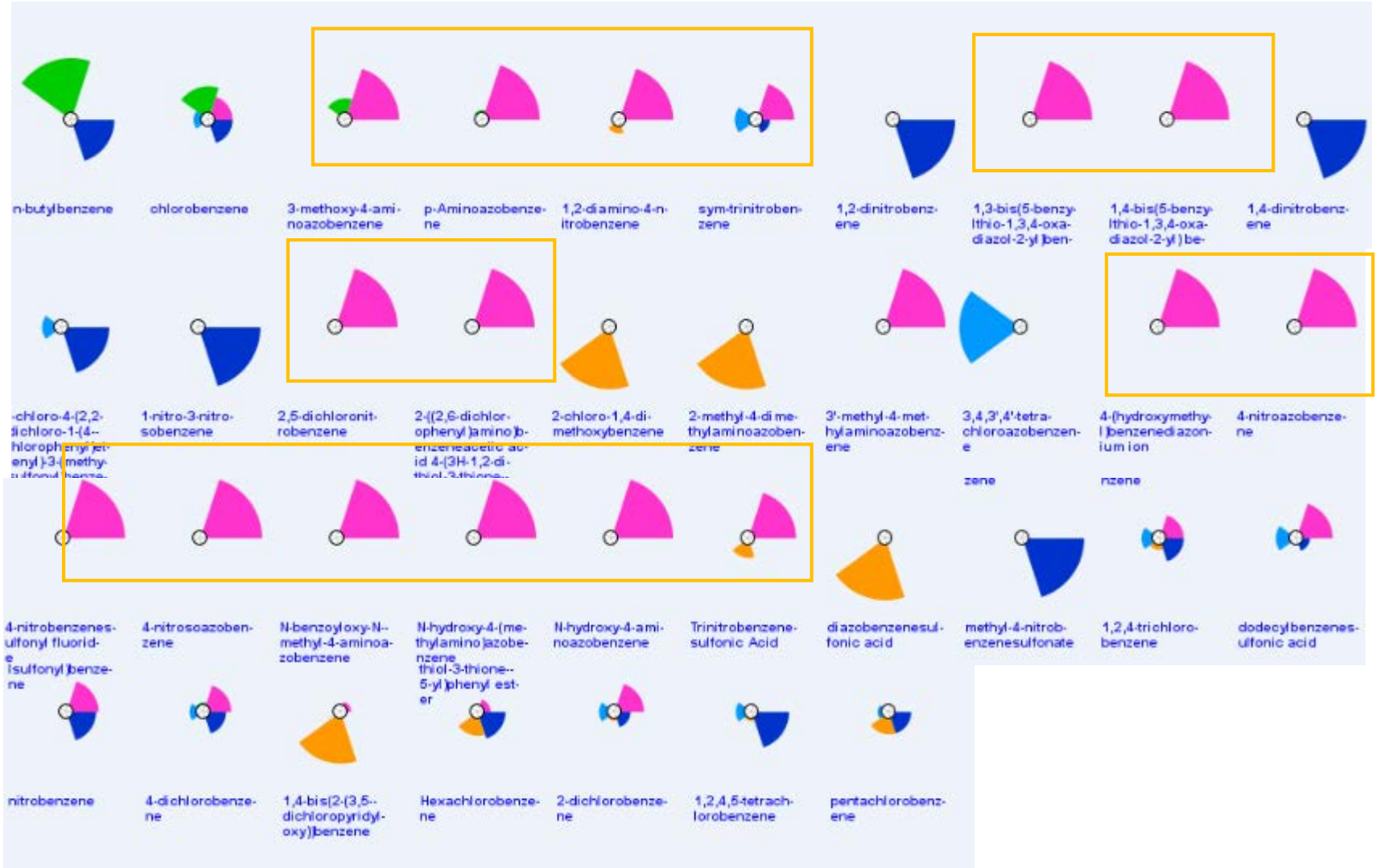
- Feed chemical names and weights into ToxPI software
- Normalize
- Produce Literature ToxPIs (LitToxPIs)

Reif DM, Martin MT, Tan SW, et al. Endocrine Profiling and Prioritization of Environmental Chemicals Using ToxCast Data. *Environmental Health Perspectives*. 2010;118(12):1714-1720. doi:10.1289/ehp.1002180.

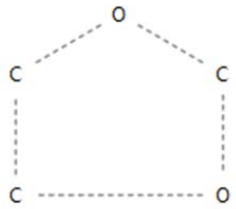
LitToxPIs for benzenes

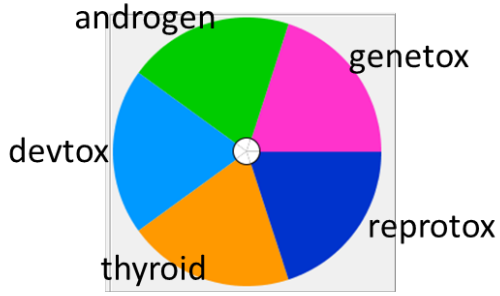


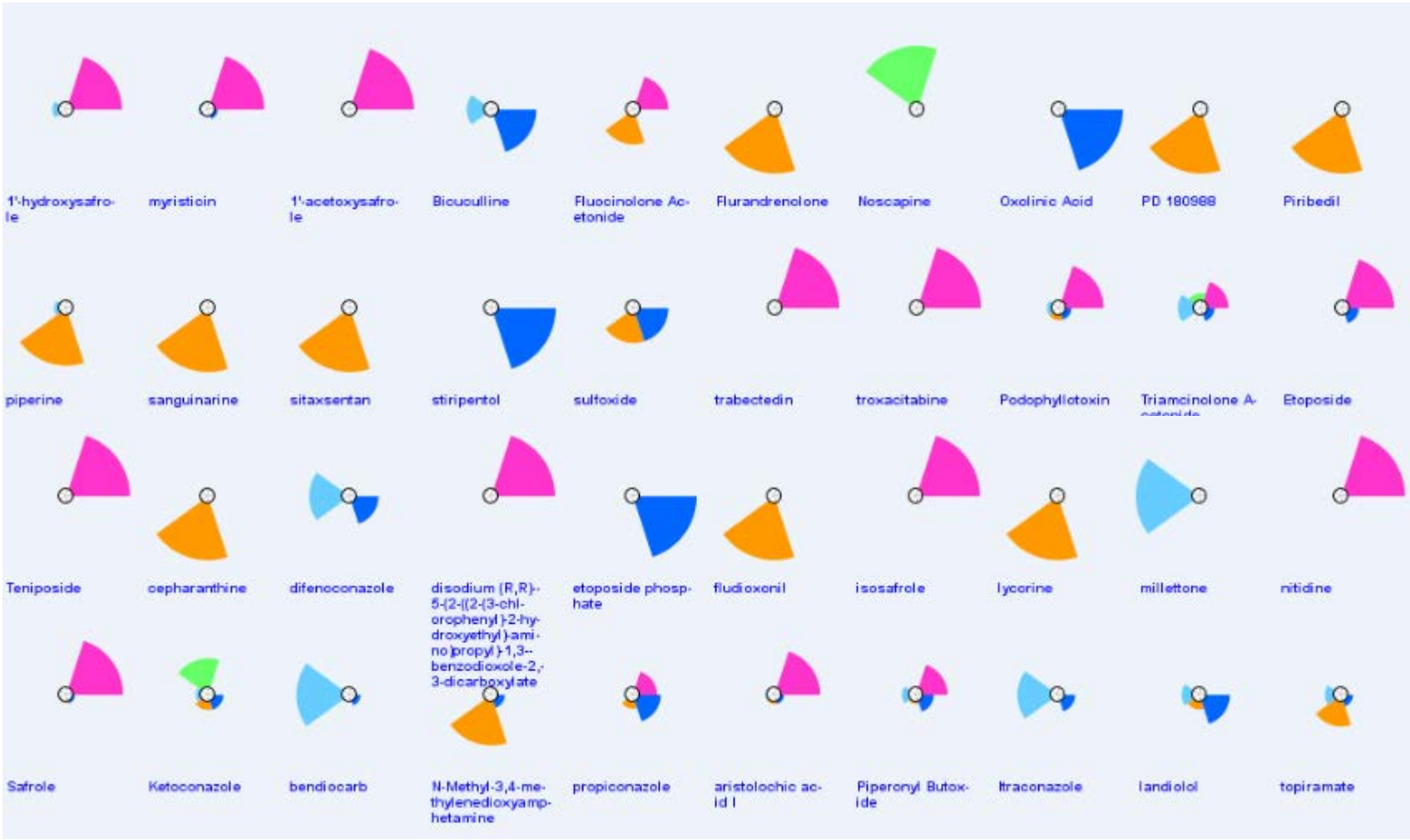
- What we can tell from this visualization?

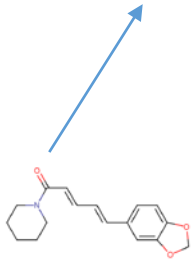


LitToxPIs for ToxPrint chemotype ring:hetero_[5]_O_dioxolane_(1_3-)

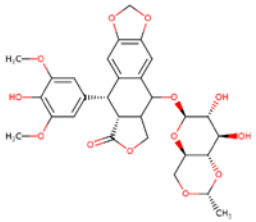








piperine

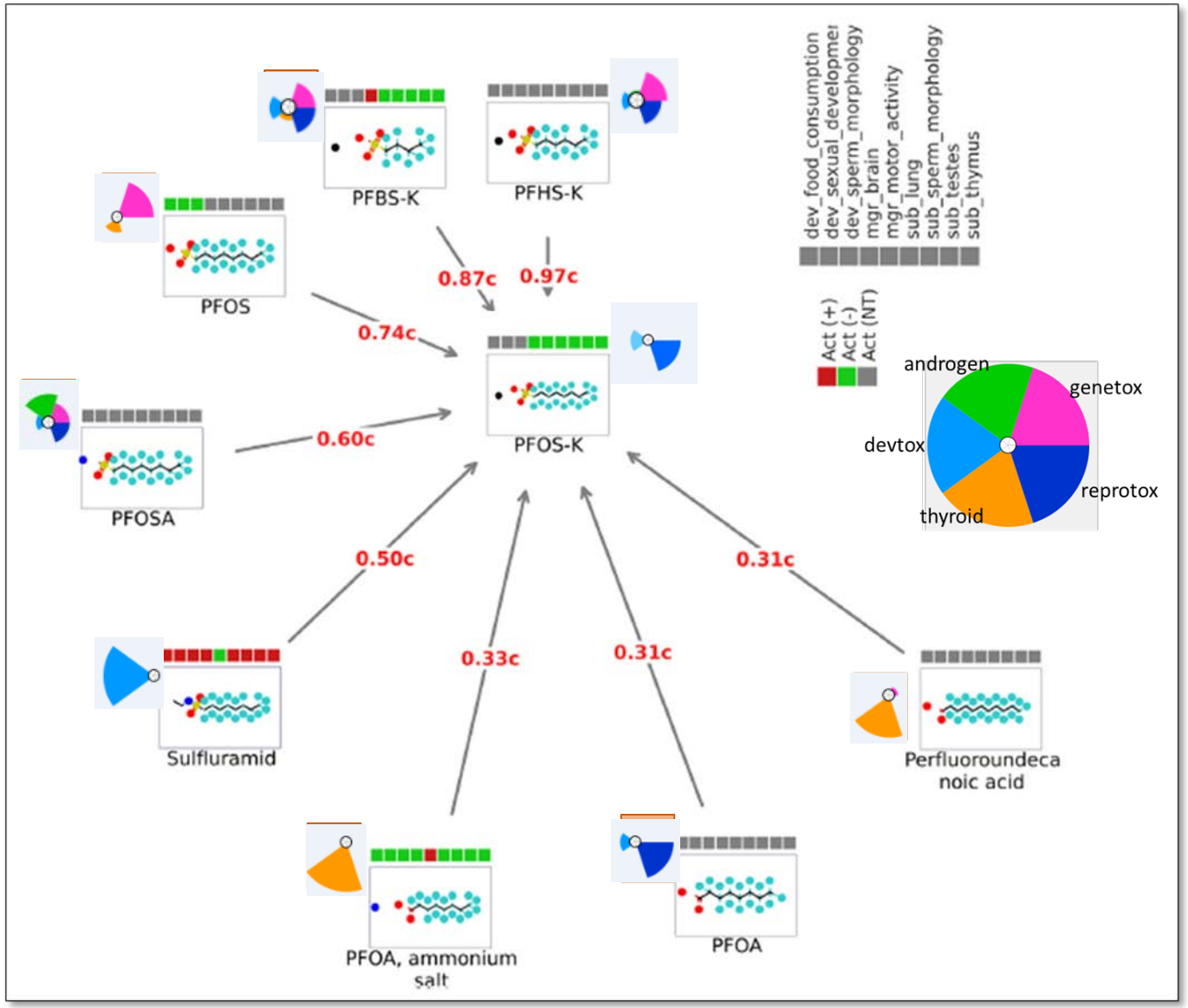


etoposide

Next steps

- Visually integrate the LitToxPI signature with the ToxRefDB signature
- Continue to refine and enhance the gathering, organizing, condensing, weighting, and normalizing steps

Illustration only.



Conclusion

- Literature mining algorithms and processes can condense complex unstructured literature into quantitative biological descriptors or signatures.
- This literature mined data can be used to inform and enhance chemical read-across.

Acknowledgements

- Grace Patlewicz
- Thomas Knudsen
- Kevin Crofton
- NCCT Team

Thank you!

Contact: baker.nancy@epa.gov