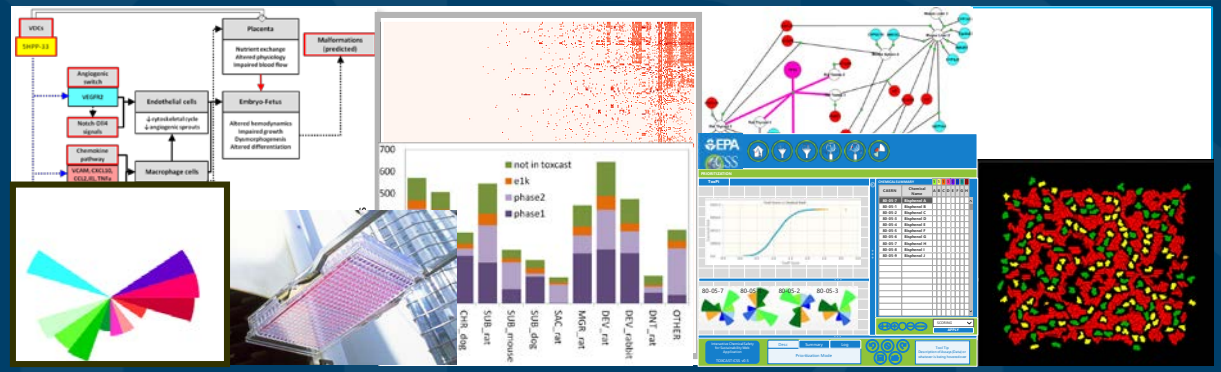
Exploiting high-throughput literature mining to support read-across predictions of toxicity



Nancy Baker*, Thomas Knudsen, Kevin Crofton, Grace Patlewicz,

National Center for Computational Toxicology
***Leidos, RTP, NC USA**

Conflict of Interest Statement

No conflict of interest declared.

Disclaimer:

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA

Outline

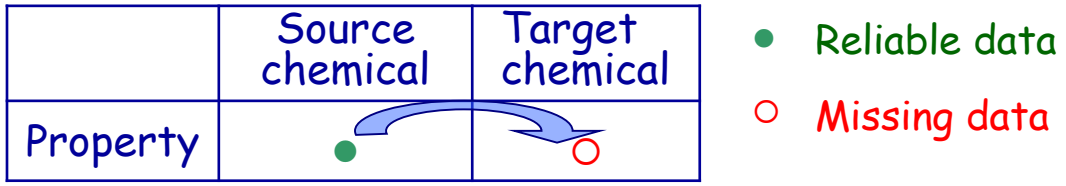
- Background and Definitions
- Workflow for category development and read-across
- Data gap analysis & challenges of data availability
- Feasibility of exploiting literature information in read-across
- Case study using skin sensitisation
- Summary

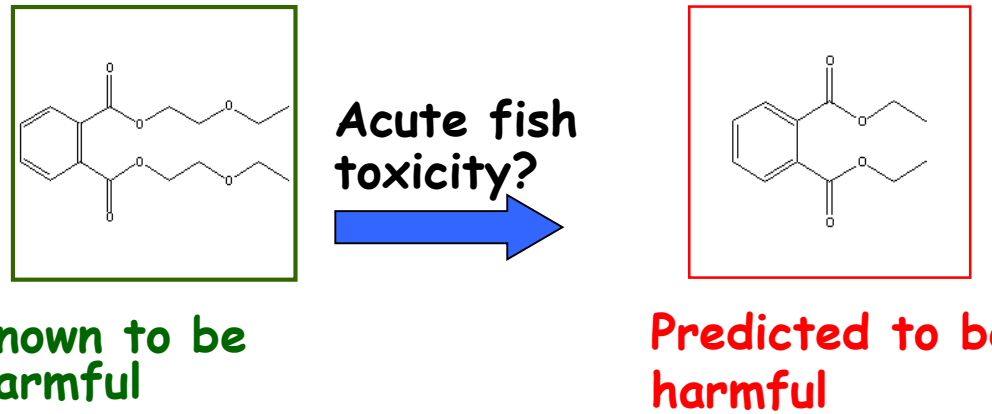
Background & definitions

- Read-across describes one of the data gap filling techniques used within analogue and category approaches
- "Analogue approach" refers to grouping based on a very limited number of chemicals (e.g. target substance) + source substance)
- "Category approach" is used when grouping is based on a more extensive range of analogues (e.g. 3 or more members)

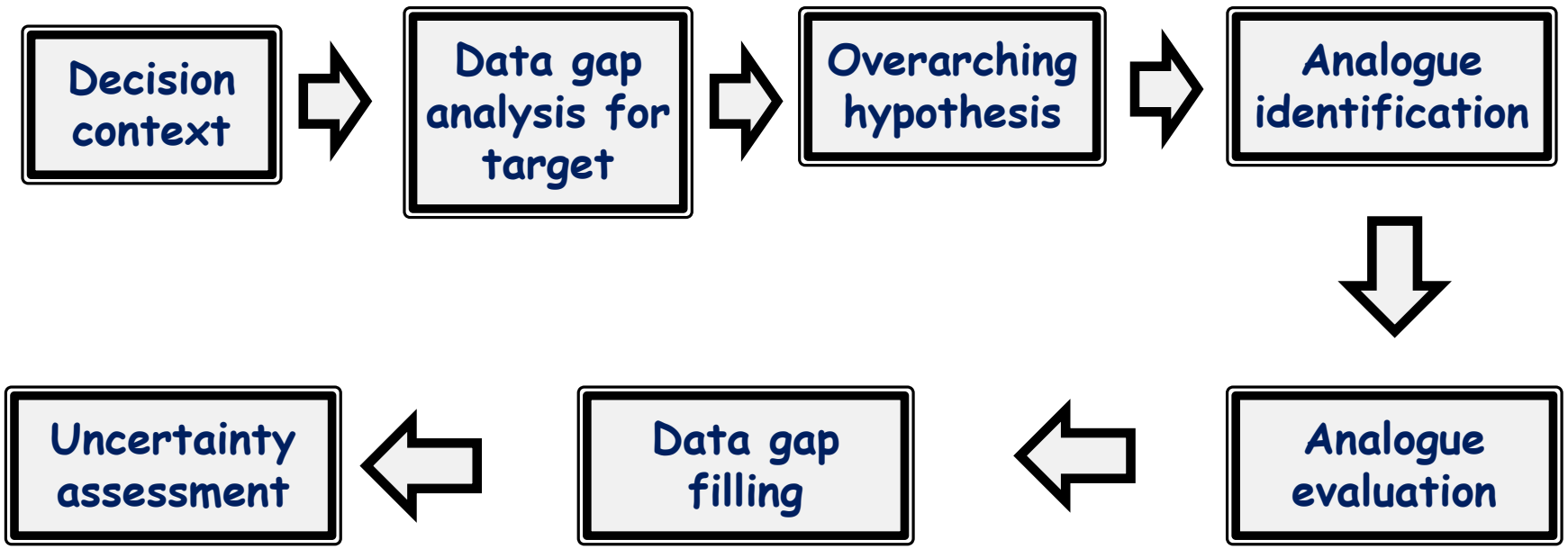
Definition: Read-across

Known information on the property of a substance (source) is used to make a prediction of the same property for another substance (target) that is considered "similar" i.e. endpoint & often study specific

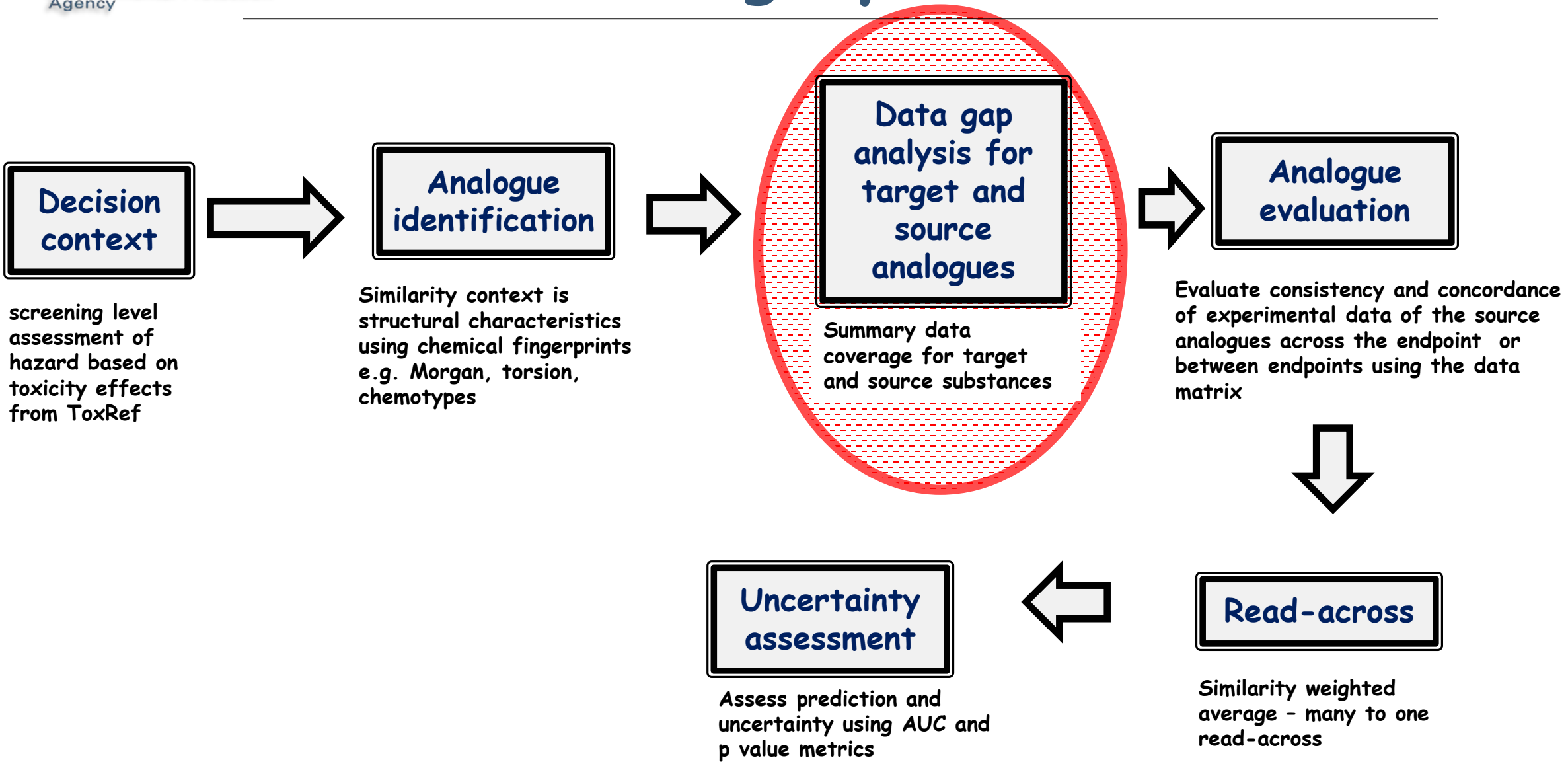




The Category Workflow



Current Category Workflow in GenRA*



*See Read-across session – Wednesday am

Data gap analysis

Data gap filling – Read-across

- Practically possible so long as a reasonable number of source analogues have been identified and evaluated with relevant data
- Typically, these data are extracted from databases such as ToxRefDB, COSMOS, ECHAChem etc – "structured data"
- However toxicity effect data may not be necessarily available for all suitable source analogues

Can literature information be helpful to inform read-across predictions? Is it feasible to identify and organise such information in a "high throughput" manner?

Literature information

- Literature information
  - Large and growing source
    - > 26 million articles; > 12 million about chemicals
    - Encompasses all sorts of toxicity effects
- Challenges
  - Literature is "unstructured"
  - Curation and validation
  - Limitations of literature mining; e.g., publication bias and granularity
  - Large and growing source – challenges for human cognition of big data

- Objective: Can we gather and condense literature information pertinent to skin sensitisation that can be useful in a read-across?

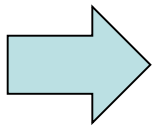
Methodology for HT literature mining to inform read-across

- Defining and organising by toxicity type
- Gathering and extracting the literature information
- Condensing and strengthening signal → signature

Defining and organising toxicity type

- Identify skin sensitisation relevant keywords, terms and categories
- Map to MeSH terms
- Score qualifying terms - to "weight" articles where a substance was closely associated with a particular study outcome

Toxicity Type
Skin sensitisation



Category	MeSH Term	Qualifier	Score
Dermatitis Contact/Atopic	Dermatitis		1
Dermatitis Contact/Atopic	Dermatitis	Chemically induced OR etiology	3
Immune Processes	Cross Reactions		1
Immune Processes	Cross Reactions	Drug effects	3
Cell	Lymphocytes		1
Cell	Lymphocytes	Drug effects	3
Chemical mediators	Cytokines		2

Gathering and extracting the literature information

Contact Dermatitis. 2006 Dec;55(6):367-8.

Allergic contact dermatitis from bisphenol-A-glycidyldimethacrylate during application of orthodontic fixed appliance.

Connolly M[1], Shaw L, Hutchinson I, Ireland AJ, Dunnill MG, Sansom JE.

Author information

PMID: 17101016 DOI: 10.1111/j.1600-0536.2006.00932.x

[PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances

Publication Types

Case Reports

MeSH Terms

Adolescent

Allergens/adverse effects*

Bisphenol A-Glycidyl Methacrylate/adverse effects*

Dental Cements/adverse effects*

Dermatitis, Allergic Contact/diagnosis*

Dermatitis, Allergic Contact/etiology

Arch Toxicol. 2011 Nov;85(11):1453-61. doi: 10.1007/s00204-010-0593-x. Epub 2010 Sep 29.

Bisphenol A-glycidyl methacrylate induces a broad spectrum of DNA damage in human lymphocytes.

Drozdz K[1], Wysokinski D, Krupa R, Wozniak K.

MeSH Terms

Bisphenol A-Glycidyl Methacrylate/toxicity*

Cell Cycle/drug effects

Cell Line, Tumor

Comet Assay

DNA Breaks, Double-Stranded/drug effects*

DNA Repair

DNA Repair Enzymes/metabolism

Humans

Lymphocytes/cytology

Lymphocytes/drug effects*

Condensing the literature into a signature

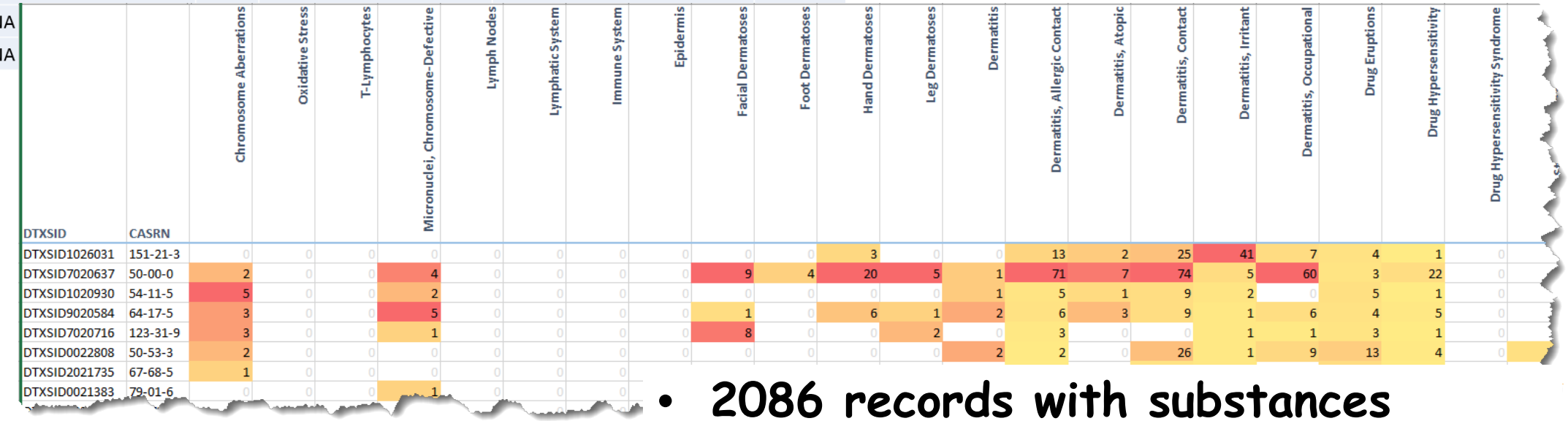
Example : Bisphenol A-Glycidyl Methacrylate

MeSH heading	Score	ToxType	Category	Category Score
Contact Dermatitis	2	Skin sensitisation	Dermatitis contact / atopic	
Dermatoses	2	Skin sensitisation	Dermatitis contact / atopic	4
Lymphocyte	3	Skin sensitisation	Cell	3
Cross Reactions	1	Skin sensitisation	Immune Processes	1
DNA Repair	2	GeneTox	DNA Damage / Repair	2
DNA / drug effects	1	GeneTox	Genetic Structures	1

Condensing the literature into a signature

From one substance to many substances…

MeSH heading	Score	ToxType	Category	Category Score
Contact Dermatitis	2	Skin sensitization	Dermatitis contact / atopic	
Dermatoses	2	Skin sensitization	Dermatitis contact / atopic	4
Lymphocyte	3	Skin sensitization	Cell	3
Cross Reactions	1	Skin sensitization	Immune Processes	2
DNA				
DNA				



DTXSID	CASRN	Chromosome Aberrations	Oxidative Stress	T-Lymphocytes	Micronuclei, Chromosome-Defective	Lymph Nodes	Lymphatic System	Immune System	Epidermis	Facial Dermatoses	Foot Dermatoses	Hand Dermatoses	Leg Dermatoses	Dermatitis	Dermatitis, Allergic Contact	Dermatitis, Atopic	Dermatitis, Contact	Dermatitis, Irritant	Dermatitis, Occupational	Drug Eruptions	Drug Hypersensitivity	Drug Hypersensitivity Syndrome
DTXSID1026031	151-21-3	0	0	0	0	0	0	0	0	0	0	3	0	0	13	2	25	41	7	4	1	0
DTXSID7020637	50-00-0	2	0	0	4	0	0	0	0	9	4	20	5	1	71	7	74	5	60	3	22	0
DTXSID1020930	54-11-5	5	0	0	2	0	0	0	0	0	0	0	0	1	5	1	9	2	0	5	1	0
DTXSID9020584	64-17-5	3	0	0	5	0	0	0	0	1	0	6	1	2	6	3	9	1	6	4	5	0
DTXSID7020716	123-31-9	3	0	0	1	0	0	0	0	8	0	0	2	0	3	0	0	1	1	3	1	0
DTXSID0022808	50-53-3	2	0	0	0	0	0	0	0	0	0	0	0	2	2	0	26	1	9	13	4	0
DTXSID2021735	67-68-5	1	0	0	0	0	0	0														
DTXSID0021383	79-01-6	0	0	0	1	0	0	0														

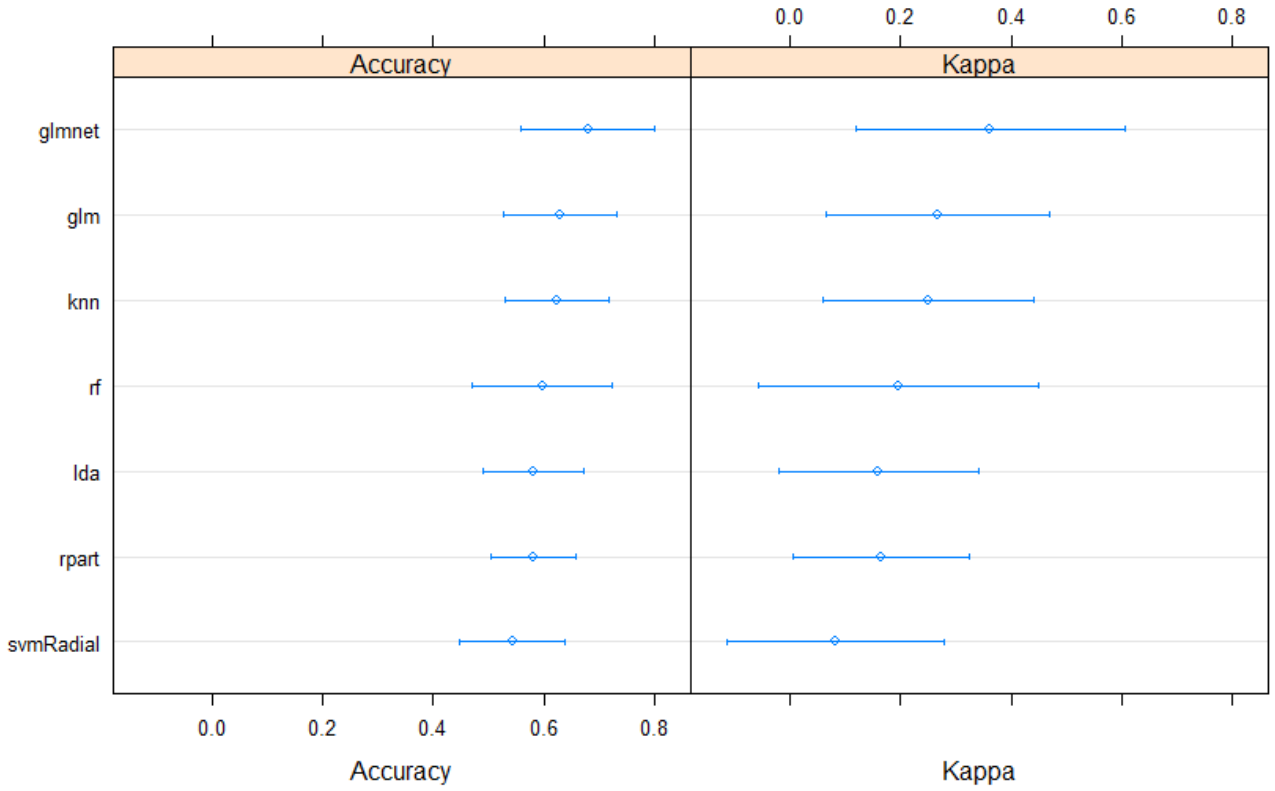
- 2086 records with substances registered within DSSTox
- 1989 defined discrete structures

Can we use this information in a read-across?

- Compiled a subset of the literature information where experimental skin sensitisation calls (i.e. from LLNA and GPMT) were known (231 substances)
- Quantitative analysis
- Are the HT literature scores correlated with LLNA/GPMT outcomes?
  - Investigated a number of machine learning approaches including logistic regression, random forest, linear discriminant analysis, SVM etc
  - Investigated whether a LitToxPi was correlated with sensitisation outcomes
- Qualitative analysis
  - Investigate the utility of the literature information to support existing experimental data for skin sensitisation to substantiate a read-across prediction

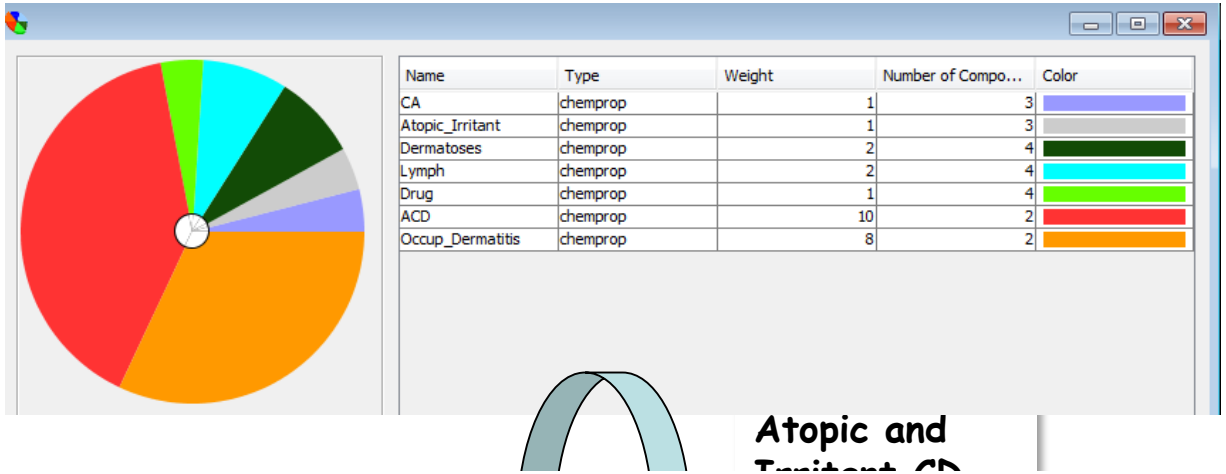
Quantitative analysis

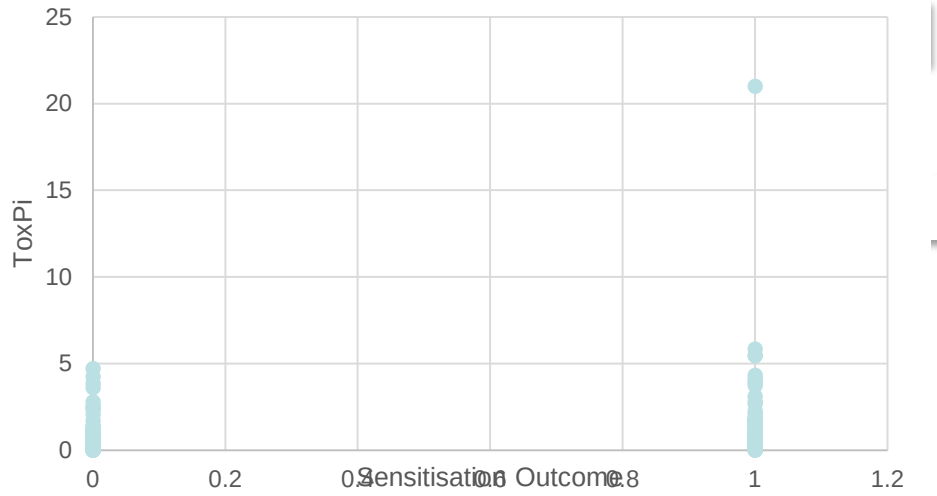
- Machine learning approaches



Accuracy ~60% or lower

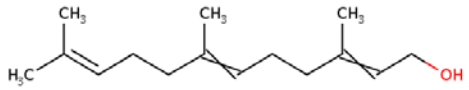
LitToxPi





Qualitative Read-across – Structural similarity

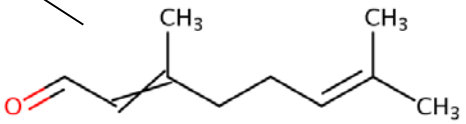
Tanimoto similarity



Farnesol

Known Sensitiser

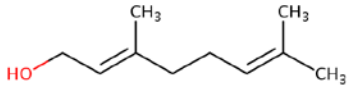
0.69



citral

Known Sensitiser

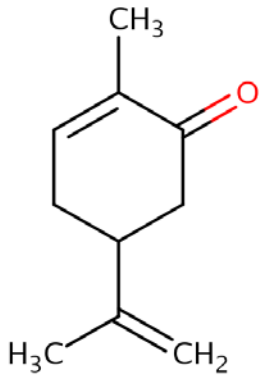
0.69



geraniol

Known Sensitiser

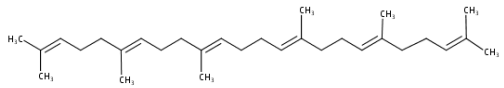
0.60



carvone

Known Sensitiser

0.65

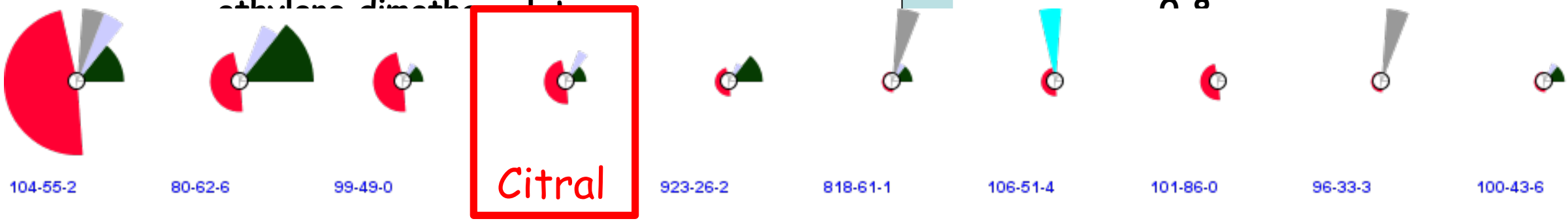


Squalene

Non-Sensitiser

- All except Squalene have a literature score>3 for ACD
- Corroborates known experimental skin sensitisation outcomes

Qualitative Read-across – Rxn mech domains



chemical_source_sid	ToxPi value
diethyl maleate	0
methyl heptine carbonate	0.588235294
quinone	2.176470588
2-hydroxyethyl acrylate	2.388235294
2-hydroxypropyl methacrylate	2.576470588
citral	3.352941176
carvone	4.329411765
Methylmethacrylate	7.329411765
cinnamic aldehyde	13.8

acceptors

Target substance

Only non-sensitiser

Summary

- Read-across is a popular data gap filling technique
- Relies upon structured information for source analogues of interest
- In the absence of such information – could literature information be exploited?
  - Proof of concept exercise to collate, structure and organise literature information relevant to skin sensitisation using the MESH terms tagged within Pubmed
  - The HT literature score was not predictive of the skin sensitisation outcome – belies the challenges of the curation, quality and biases in the literature information itself
  - However the literature information was found to be helpful in corroborating a read-across prediction for citral based on known experimental data