Application of Dempster-Shafer theory to estimate uncertainty and combine diverse sources of evidence in chemical risk assessment

James Rathman

Ohio State University

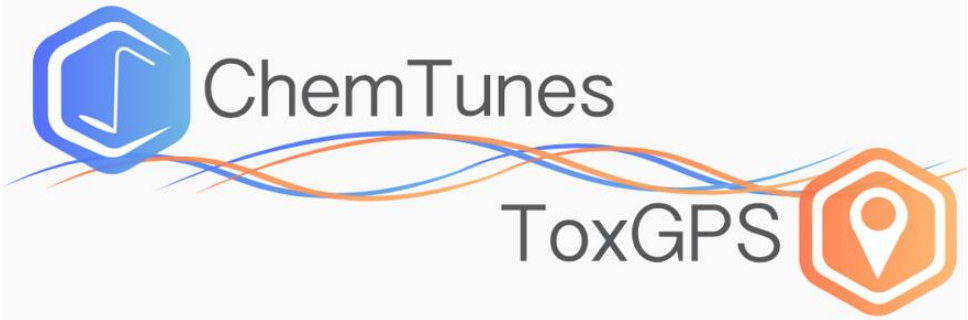
International Conference on Uncertainty in Risk Analysis

BfR

February 21-22, 2019

Chemical toxicity prediction approaches

- Quantitative structure-activity (QSAR) models
  - global and local mode-of-action models
  - descriptors
    - ToxPrint chemotypes (expert defined fragments)
    - Physicochemical properties: logP, logS, TPSA, shape descriptors, etc.
    - Quantum mechanical properties: HOMO, LUMO, heat of formation
- Structural rules
  - expert-guided knowledgebase
- Read-across
  - using data available for suitable analogs to infer toxicity of a target compound
- Weight-of-evidence outcome using Dempster Shafer Theory

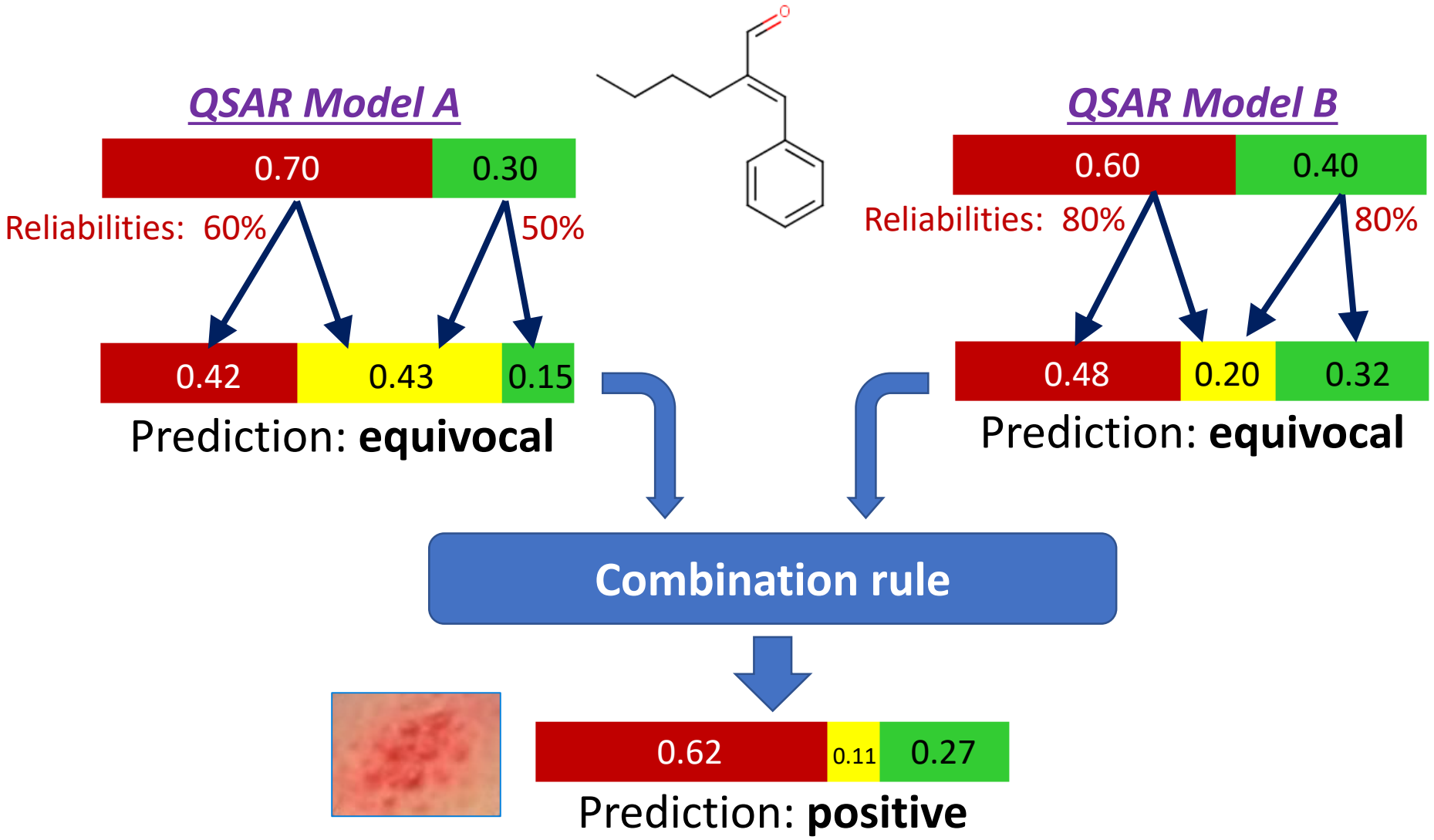


Dempster-Shafer Theory (DST)

- DST provides a rigorous approach for:
  - estimating uncertainty
  - combining multiple sources of evidence to make a decision
- Allows us to explicitly take into account:
  - reliability of quantitative structure- activity (QSAR) models
  - reliability of structural rules (“alerts”)
  - reliability of experimental results from in vitro assays and toxicity studies

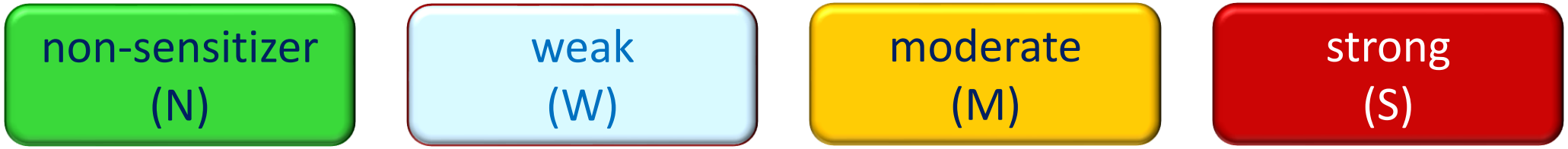
Rathman, J.F., Yang, C., Zhou, H. “Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment”, *Computational Toxicology 6*, 16-31 (**2018**)

Skin sensitization prediction

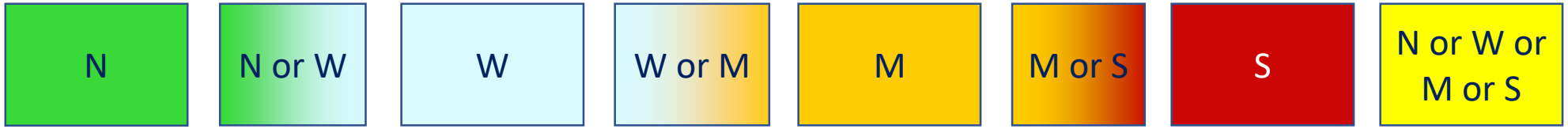


Ordinal classification

Consider a four-level classification model for skin sensitization:



The Dempster-Shafer focal elements can be defined such that the model has 8 possible prediction outcomes:



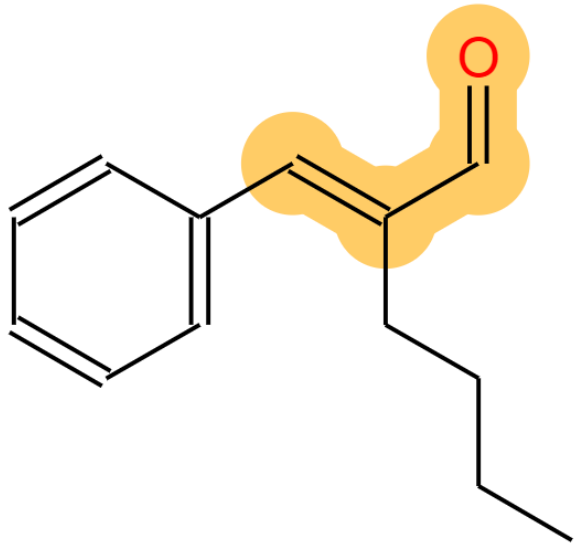
DST allows us to capture different degrees of uncertainty.

Reliability measures for QSAR classification models

- Performance statistics from model validation
  - accuracy (concordance, Matthews correlation coef)
  - sensitivity and specificity
  - positive and negative predictive values (PPV and NPV)
- Domain of applicability
- Ideally, an independent external test set should be used…
- …but for many toxicity endpoints, high-quality data suitable for building QSAR models are limited. We may then need to rely on cross-validation performance measures.

Reliability measures for structural rules

Example of a chemotype alert for skin sensitization



α,β-unsaturated ketone
(Michael acceptor)

odds ratio = 5.28

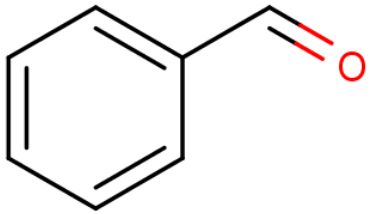
Reliability measures for toxicity studies

- ECVAM-validated methods with reliability estimates (e.g., DPRA, Keratinosens™, and h-CLAT assays for skin sensitization).
- Klimisch scoring based on assessment of how well a toxicity study conforms to internationally accepted testing guidelines.
  - 1 = reliable without restriction
  - 2 = reliable with restriction
  - 3 = not reliable
  - 4 = not assignable
- When the original study data are not available, Klimisch scores, if not provided, cannot be extracted; or, if provided, cannot be verified.

H.J. Klimisch, M. Andreae and U. Tillmann, "A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data", *Regulatory Toxicology and Pharmacology, 25*, 1–5 (1997).

Accounting for uncertainty of in vitro assays

Example: skin sensitization (LLNA) for benzaldehyde



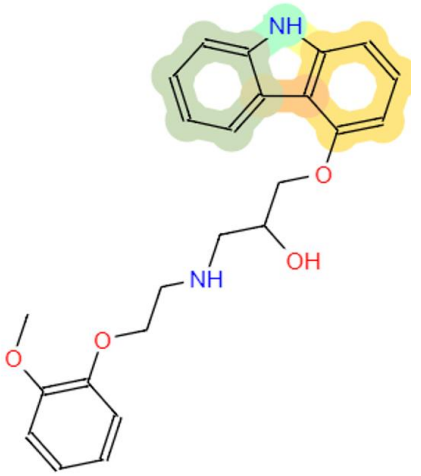
Performance metrics

In vitro assays	PPV	NPV
DPRA	0.87	0.57
KeratinoSens™	0.85	0.52
h-CLAT	0.85	0.57

Urbisch, et al. (*Reg Tox and Pharm 71*, **2015**, 337-351)

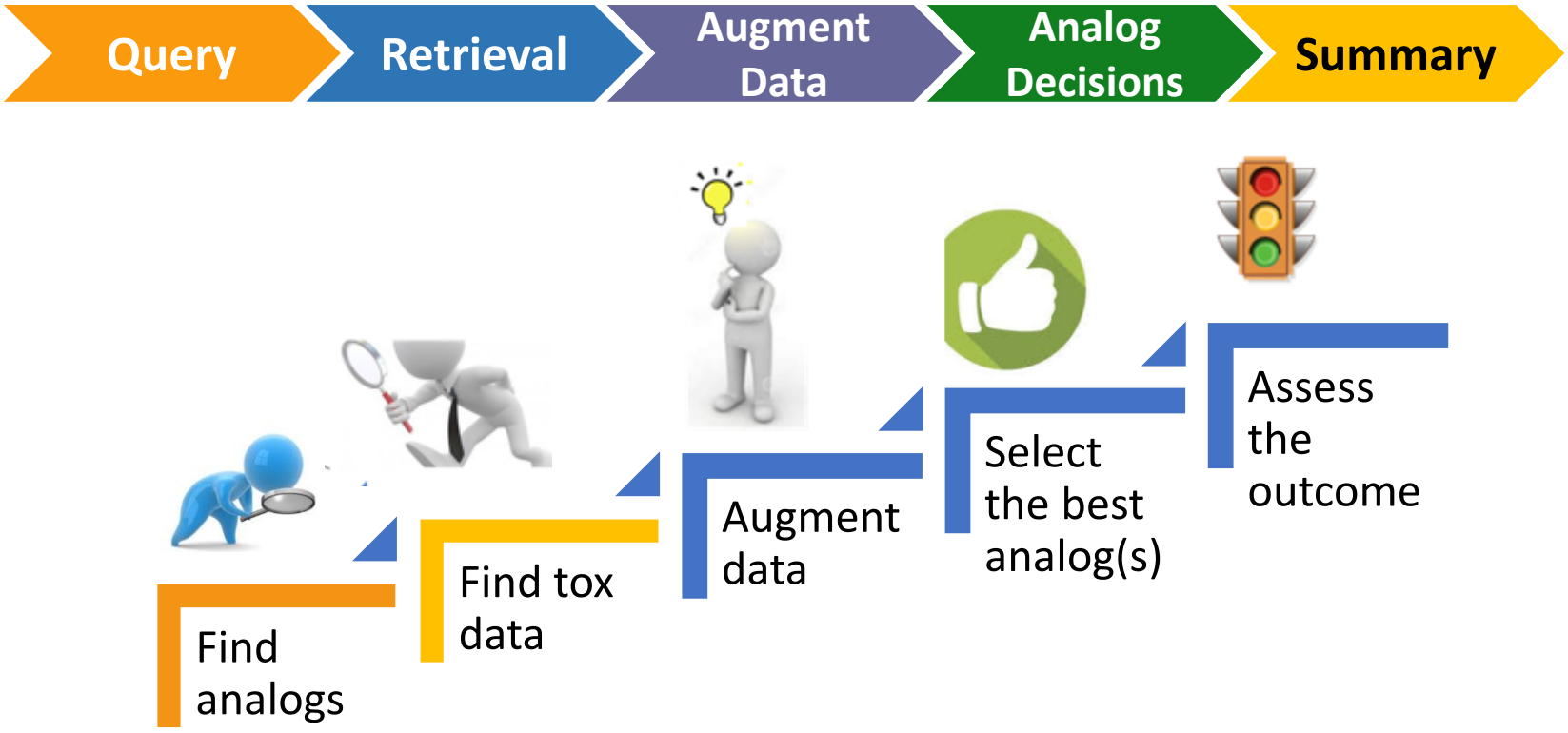
In vitro assay	Assay result	DST Probabilities	LLNA Prediction
DPRA	negative	0.57 | 0.43	negative
KeratinoSens	positive	0.15 | 0.85	positive
h-CLAT	positive	0.15 | 0.85	positive

Factors that reduce reliability



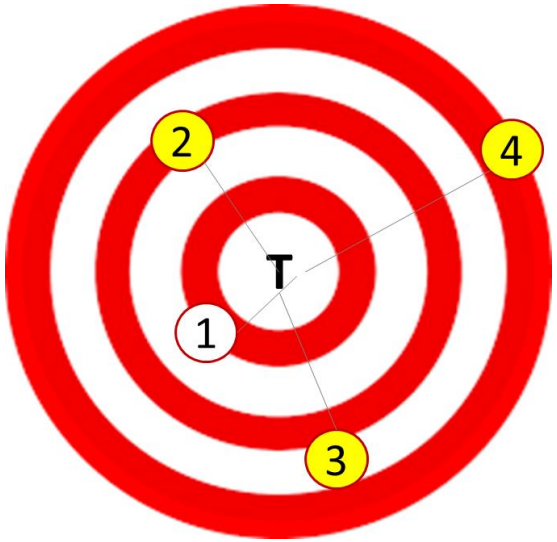
- Inaccurate chemical structures
- Chemical reactivity, metabolism
  - test material differs from the active entity
- Problematic toxicity study results
  - secondary or tertiary data sources (e.g. databases, safety assessment reports) may be not be precise or exhaustive, or may introduce mistakes
  - lack of information on guideline (GLP-compliant?), certain study design parameters (route of exposure, doses tested, etc.), or critical effects
- Inconsistent calls for a given toxicity endpoint
  - compound level (multiple studies with different calls)
  - study level (same study with different calls depending on the regulatory body/organization responsible for the call)
- Limited or unspecified domain of applicability

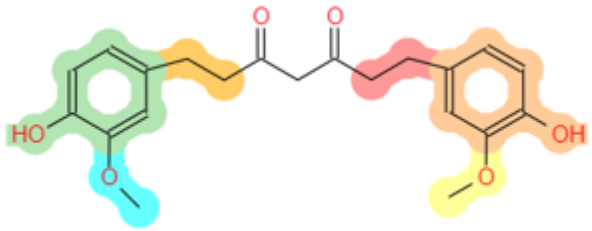
Generic read-across workflow



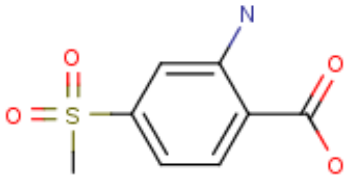
Collecting evidence for read-across

- Find analogs and evaluate analog quality based on
  - structure similarity
  - property similarity
- Apply chemotype profilers for relevant biology
  - DNA binders
  - protein binders
  - metabolic rules
- Consider metabolism
  - metabolite generation
  - metabolic similarity
- Find tox data for analogs

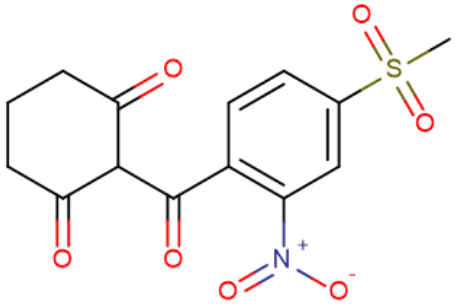




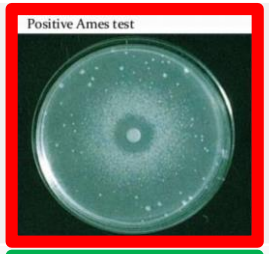
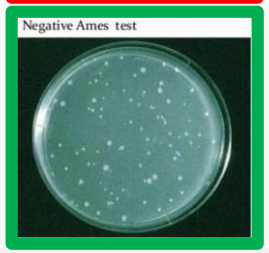
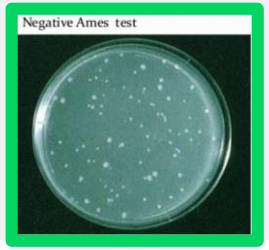
Read-across example using Ames results for a single analog



target: metabolite of Mesotrione



analog: Mesotrione

Ames Assay Result	Study Reliability	Analog Quality	Probability Bar
	0.50	0.62	
	0.95		
	0.80		

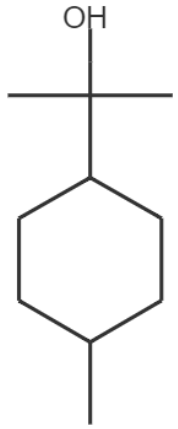
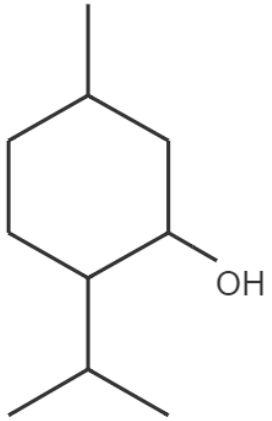
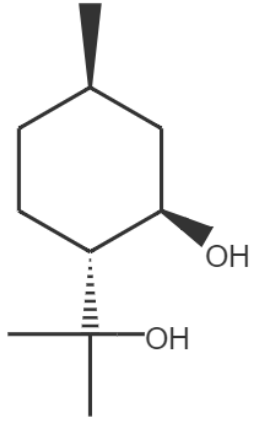
DST Combination

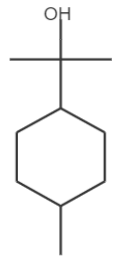
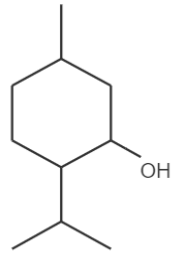
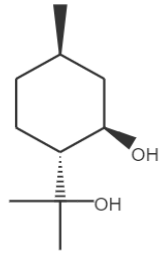
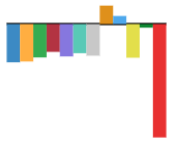
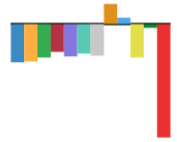
Read-Across Outcome
NEGATIVE

Ames assay images: www.mun.ca/biology/scarr/4241_Ames_test_reversion.html

Read-across example with multiple analogs

Read-across for repeated-dose toxicity of dihydro-α-terpineol from menthol and menthanediol:

dihydro-α-terpineol	menthol	menthanediol
		

Compound Summary	target	analog 1	analog 2
CMS ID			
Data Summary			
Studies	0	5	1
Fingerprints			
RDKit MolFingerprint			
Tanimoto		0.8	0.76
ToxPrint Fingerprint			
Tanimoto		0.5	0.71
Skyline Profiles			
Skyline Terpineol			
Skyline			
Pearson correlation coefficient		1	0.97
Analogue Quality		**0.74**	**0.81**

Short-term RDT Study	target	analog 1	analog 2
Description		Rat, oral-gavage, 28 days	
Outcome		LOEL = 200 mg/kg BW/day, Liver	
Reliability		Low (by RepDose)	
reliability score: Reliability		0.5 positive ×	
Subchronic RDT Study			
Description		Rat, oral, 91 days	
Outcome		NOEL = 50 mg/kg BW/day, Organ Weight	
Reliability		Low	
reliability score: Reliability		0.5 positive ×	
Chronic RDT Study			
Description		Rat, oral, 730 days	
Outcome		NOAEL = 750 mg/kg BW/day, Body Weight	
Reliability		High	
reliability score: Reliability		0.9 negative ×	
DART Study			
Description			Rat, DART
Outcome			NOAEL = 400 mg/kg BW/day, Pub Weight
Reliability			Medium
reliability score: Reliability			0.7 negative ×

Predicted Toxicity	target	analog 1	analog 2
Cleft Palate	negative		
Probability Bar			
Oral hDILI			
Call	negative		
Probability Bar			
Analogue Quality		0.74	0.81
TIER 1 (Analogue+Exp)	 negative		
TIER 2 (Analogue+Exp+In silico)	 negative		

Real-world experience

- Our goal is to help experts in regulatory bodies and industry make good decisions. They want methods that are
  - transparent
  - interpretable and mechanistic
  - as simple as possible
- They are often uncomfortable reporting decisions with any appreciable uncertainty, or if there are conflicting pieces of evidence.
- They want the decision-making process to be interactive, but may be unsure about how to select good analogs, choose evidence sources, or specify reliabilities.
- Experts looking at the same evidence will not always agree, but DST-based approaches can help identify *why* they disagree.

Jim Rathman
james.rathman@mn-am.com
rathman.1@osu.edu

Chihae Yang
chihae.yang@mn-am.com

Aleksandra Mostrag
aleksandra.mostrag@mn-am.com

Mark Cronin
M.T.Cronin@ljmu.ac.uk