Identifying, Estimating, and Communicating Uncertainty in *In Silico* Approaches to Chemical Safety Assessment

Jim Rathman, PhD

Chemical & Biomolecular Engineering
The Ohio State University, Columbus, OH
rathman.1@osu.edu

MN-AM, Nürnberg, Germany, and Columbus, OH
james.rathman@mn-am.com

Conflict of Interest Statement

- No conflicts of interest to declare

Uncertainty

- Types of uncertainty
  - Aleatory (statistical, “random error”)
  - Epistemic (knowledge limits, “ignorance”)

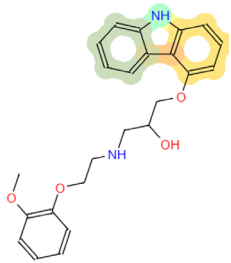
- Epistemic uncertainty may be reduced by improving models through incorporation of relevant mechanistic knowledge from chemistry and biology.
- Although we routinely deal with uncertainty in everyday life, many people do not expect or accept uncertainty when it come to scientific outcomes.

In Silico Approaches to Chemical Toxicity Prediction

- 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
- Read-across
- Weight-of-evidence combination of information

Sources of Uncertainty in QSAR and Rule-based Approaches (Cronin et al., 2019)

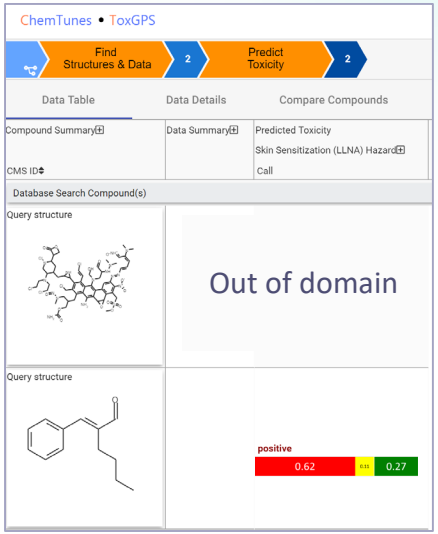
- Chemical structures and descriptors
  - Inaccurate structures
  - Chemical reactivity, metabolism (computational form differs from active form)



- Biological data
  - Secondary or tertiary data sources (e.g., databases, safety assessment reports) may not be precise or exhaustive, or may introduce mistakes
  - Lack of information on guideline, study design parameters, or critical effects
  - Inconsistent calls (e.g., same study with different calls depending on the regulatory body/organization responsible for the call)
  - Limited knowledge of mechanism of action
  - Relevance of data to endpoint being modeled

- Descriptors
  - Structural features used to fingerprint molecules
  - Coverage
  - Experimental/calculated properties
  - Relevance of descriptors

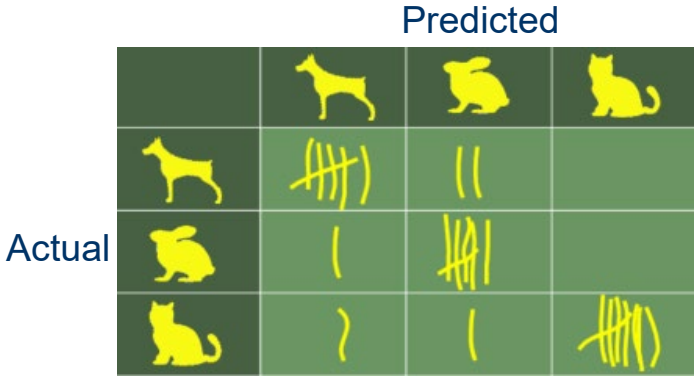
- Undefined (or poorly-defined) domain of applicability



- Some of these can be addressed by careful curation of the chemical and biological data.
- For others, we want to *quantify* uncertainty to estimate how these factors affect reliability of the prediction results.

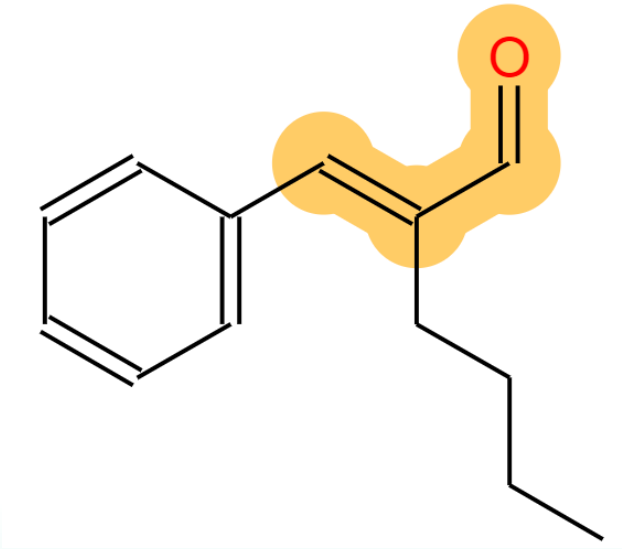
Reliability Measures for QSAR Classification Models

- Performance statistics from model validation
  - Accuracy (concordance, Matthews correlation coef)
  - Sensitivity and specificity
  - Positive and negative predictive values
- Training set and validation
  - Size and balance
  - Coverage of relevant chemical space
- Domain of applicability



Reliability Measures for Structural Rules

Example of a chemotype alert for skin sensitization



α,β-unsaturated ketone
(Michael acceptor)

odds ratio = 5.28

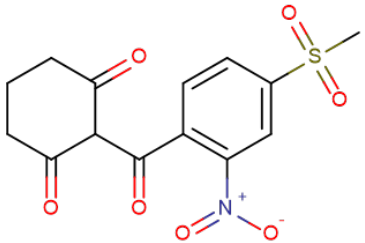
Reliability Measures for Toxicity Studies

- 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

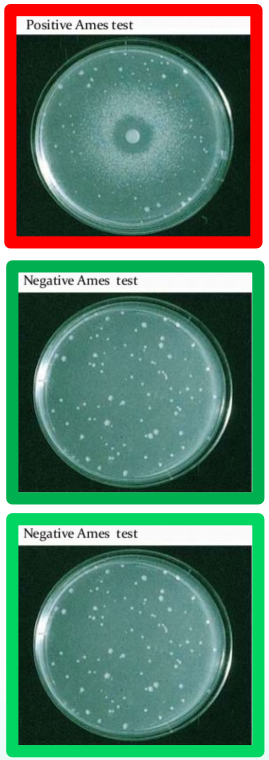
  (Klimisch et al., 1997)
- Klimisch scores can only be determined and verified if the original study data are available.

Strategies for Combining Multiple Pieces of Evidence

Bacterial reverse mutation assay (Ames test)



Mesotrione



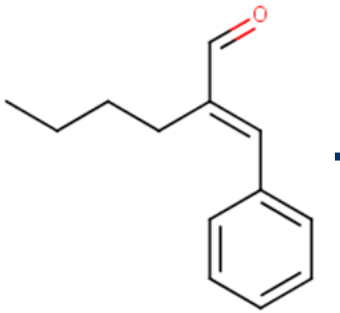
Strategy	Prediction
Majority wins (voting)	NEGATIVE
One or more (*N* or more)	POSITIVE
Probability bounds analysis	?

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

Probability Bounds Methodology (Rathman et al., 2018)
(Based on Dempster-Shafer Theory, 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

Example: Skin Sensitization Hazard Prediction



QSAR Model A

PPV = 0.60
NPV = 0.50

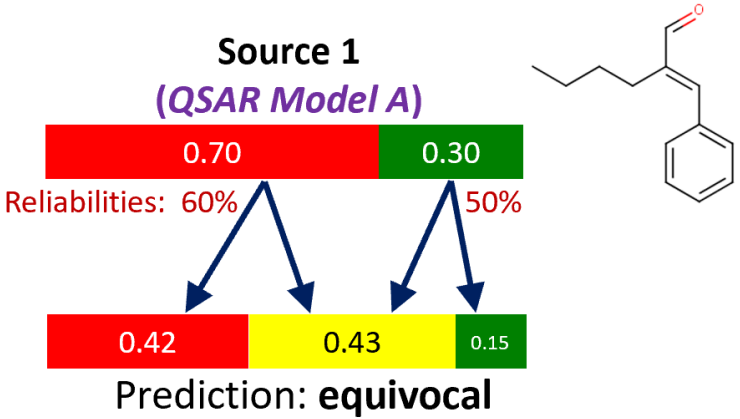
$p(POS) = 0.70$

QSAR Model B

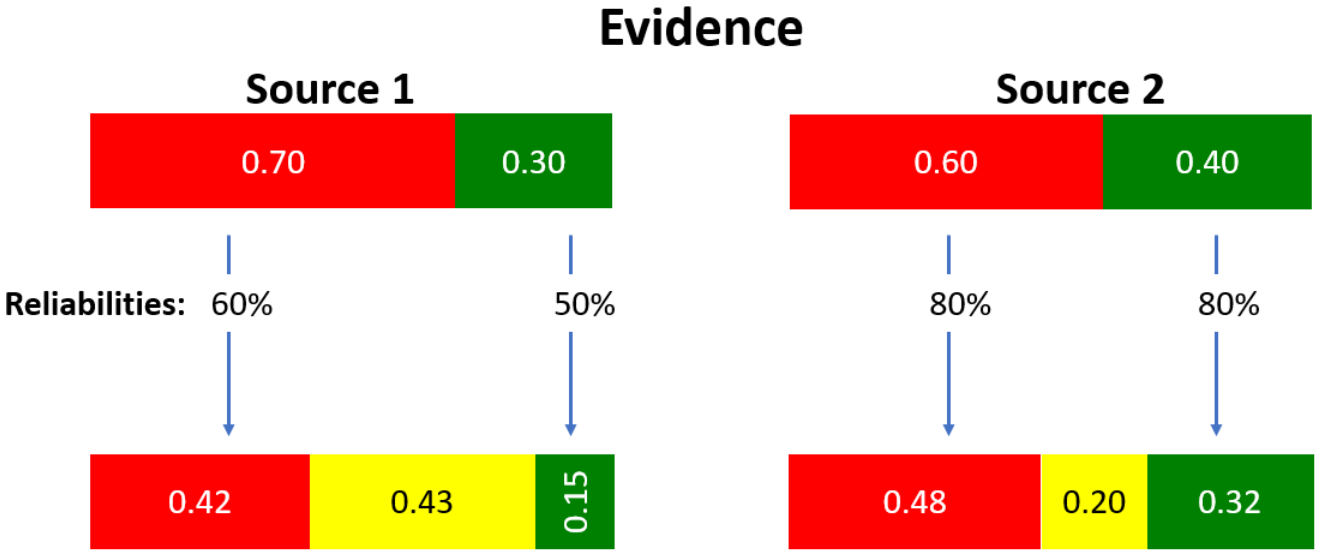
PPV = 0.80
NPV = 0.80

$p(POS) = 0.60$

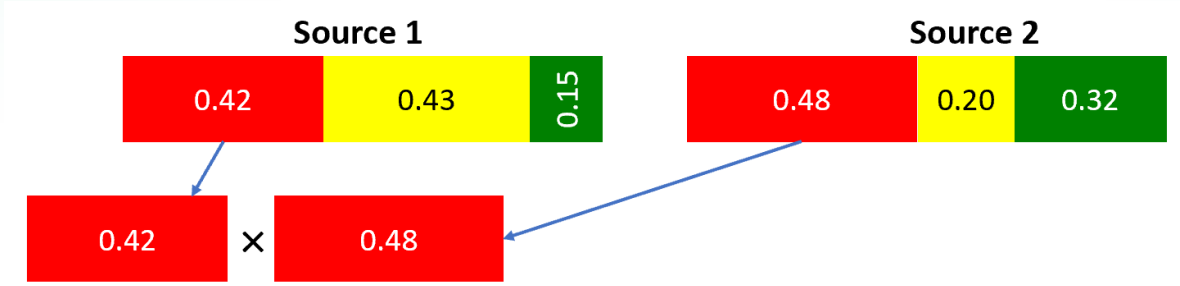
Prediction and Combination



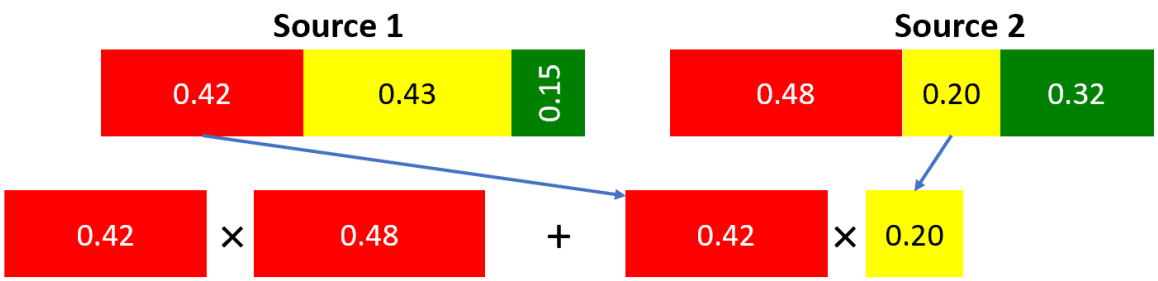
Probability Bounds Analysis of Single Sources



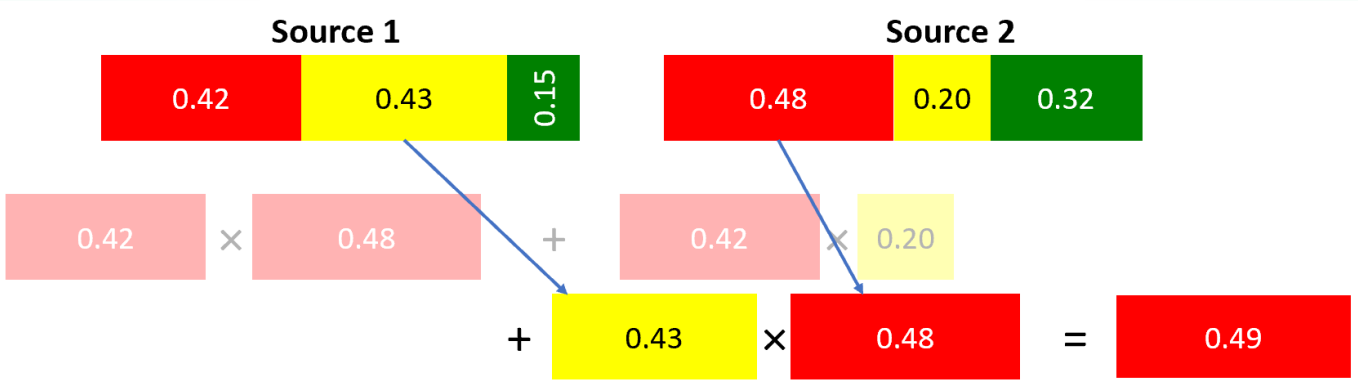
Dempster Combination Rule



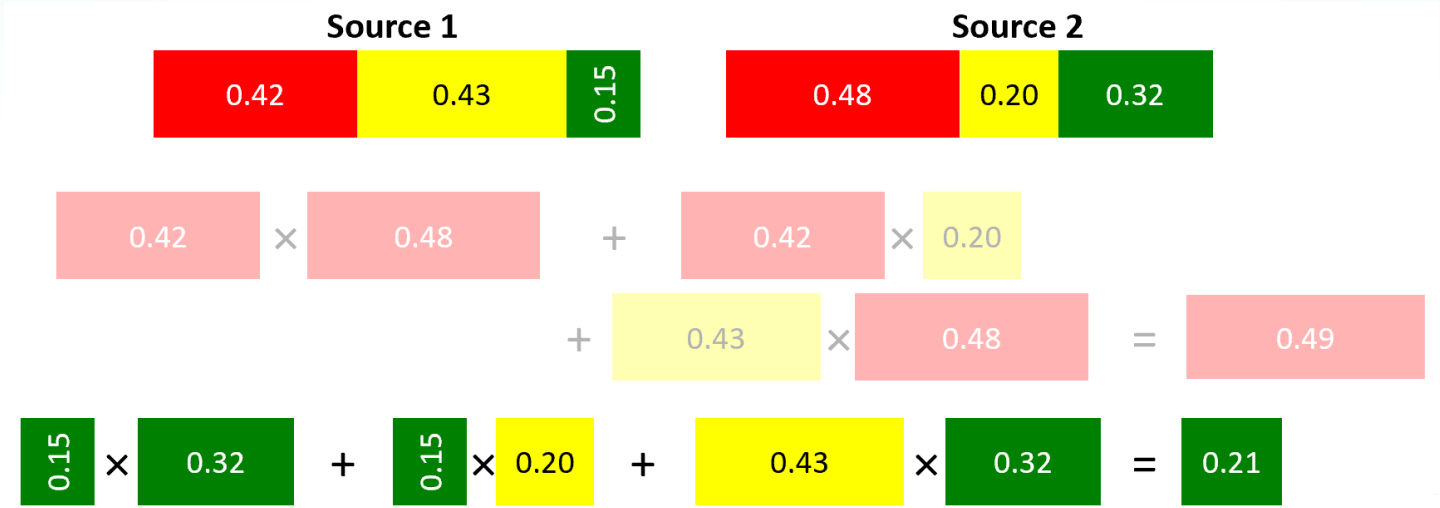
Dempster Combination Rule



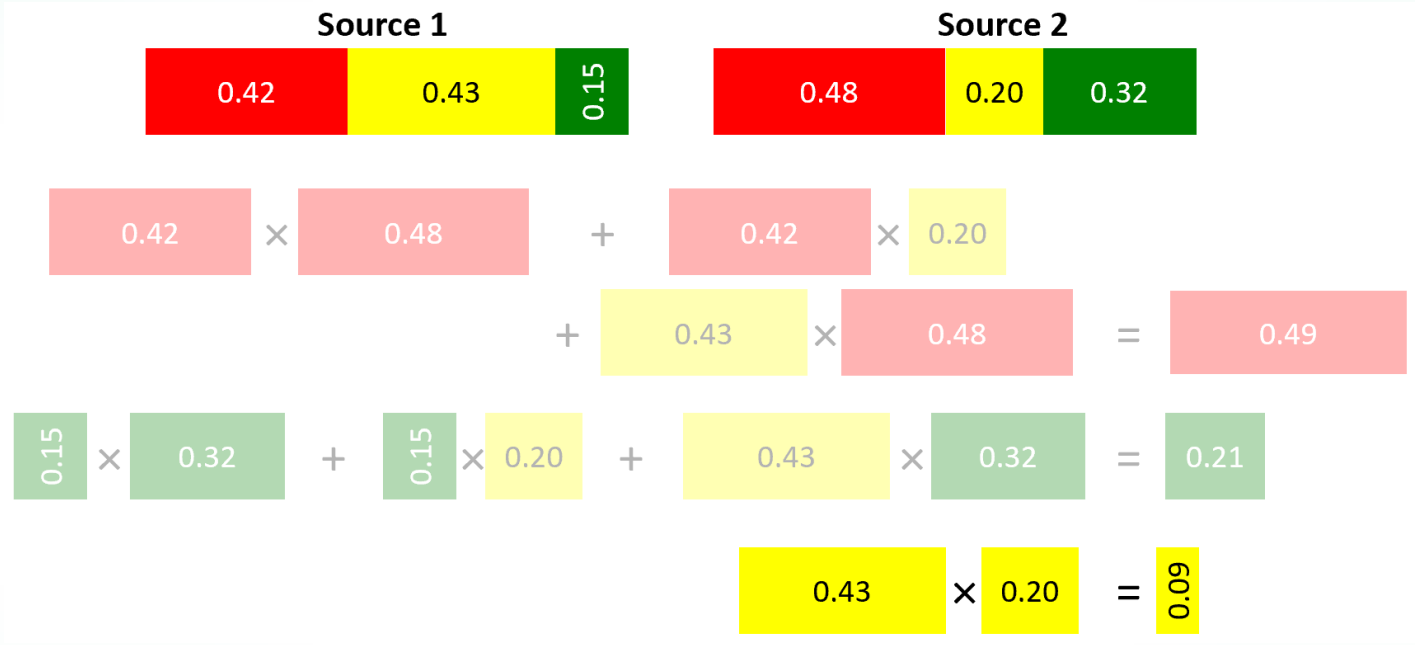
Dempster Combination Rule



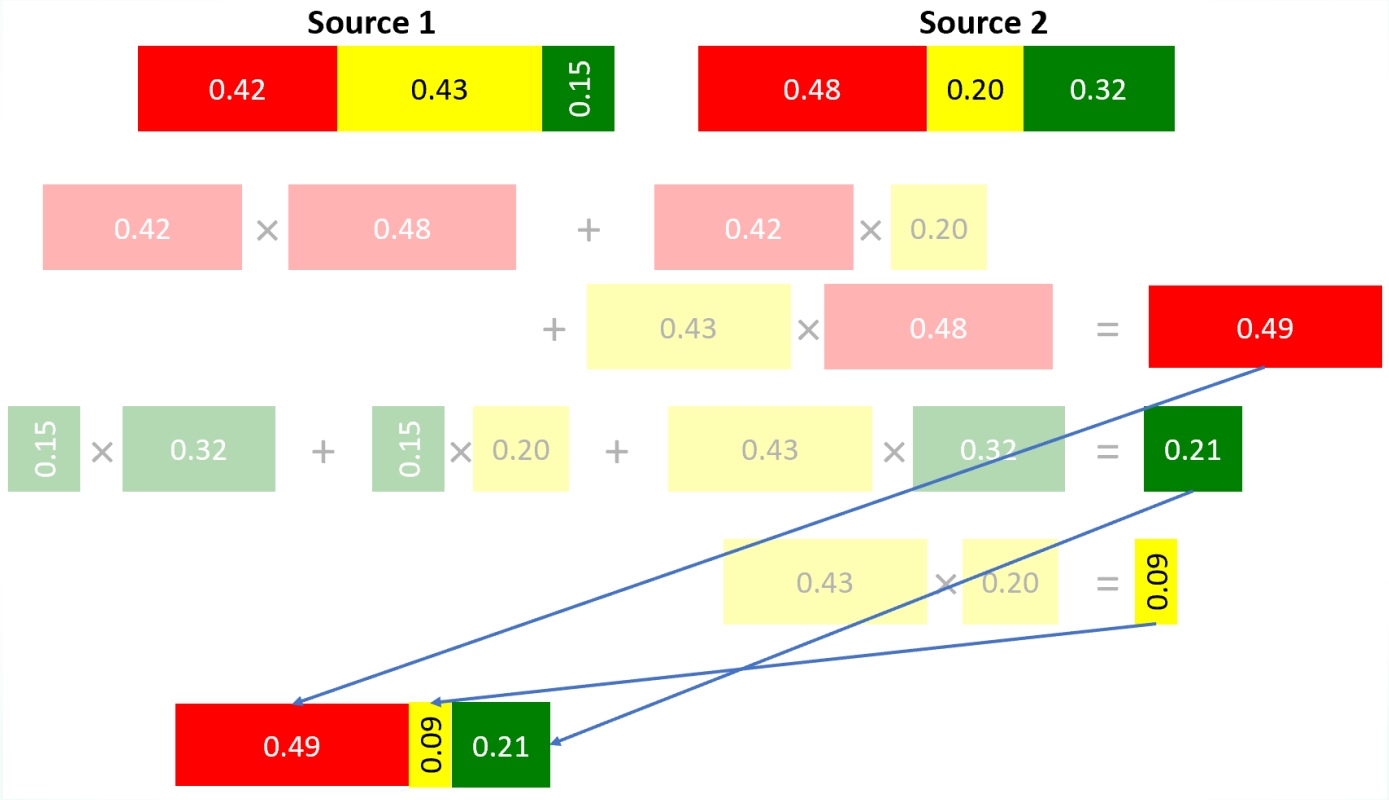
Dempster Combination Rule



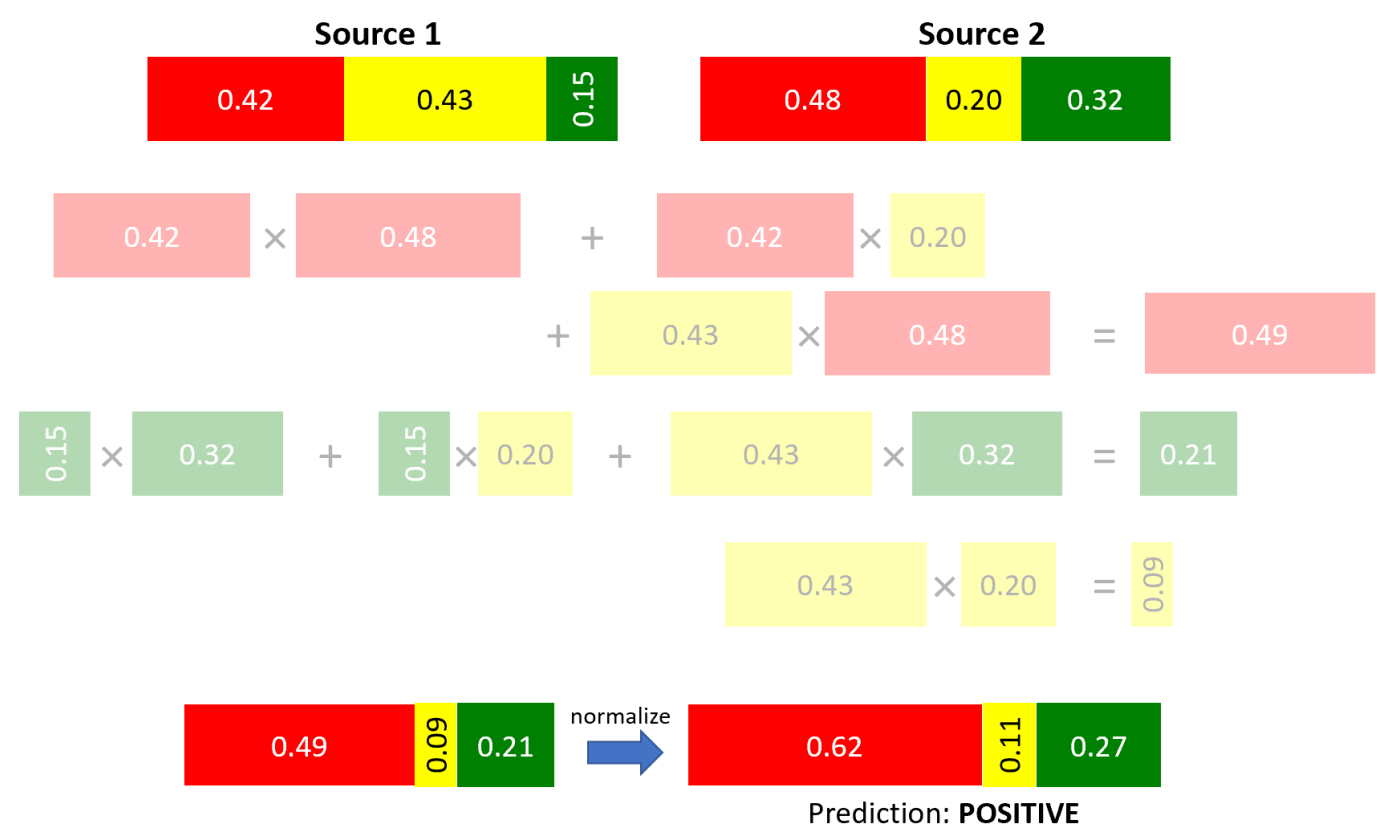
Dempster Combination Rule



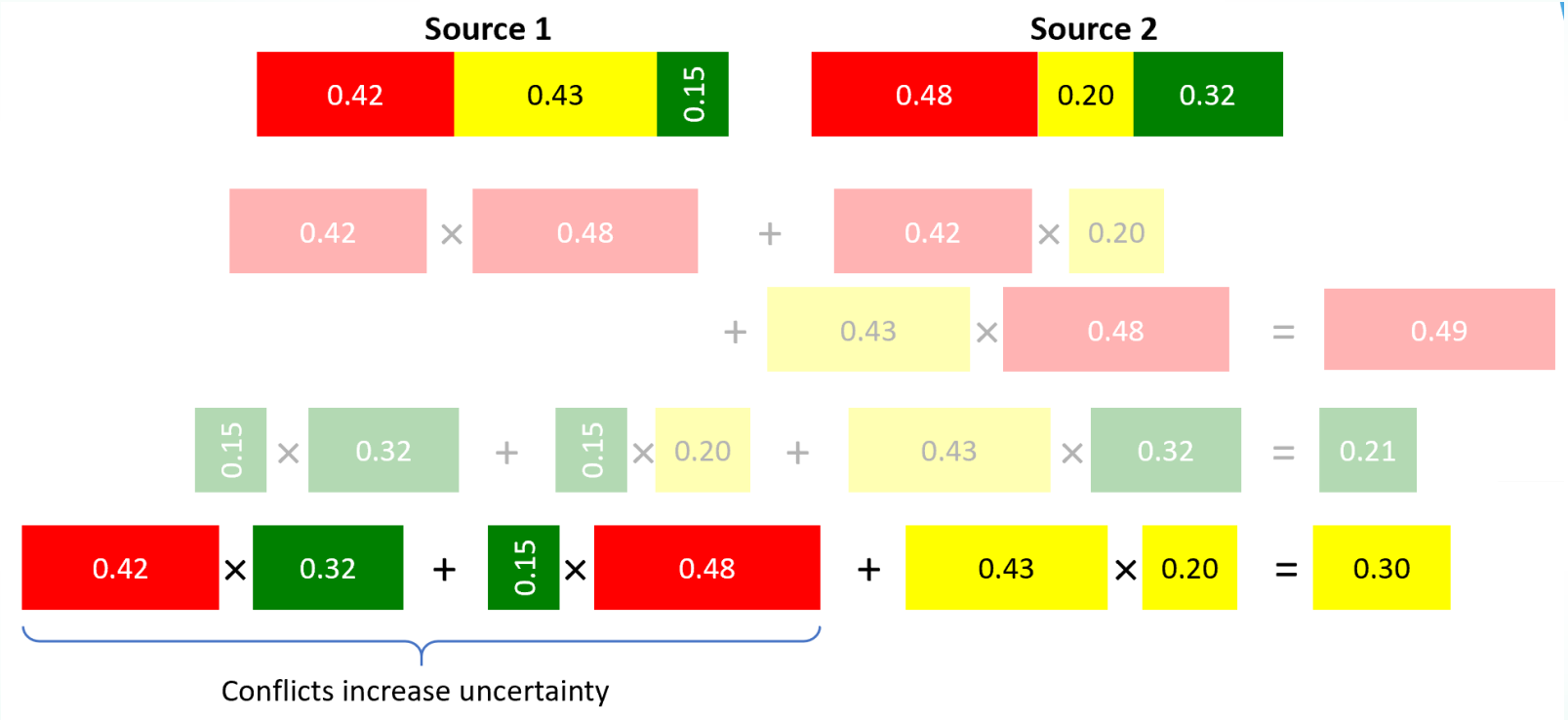
Dempster Combination Rule



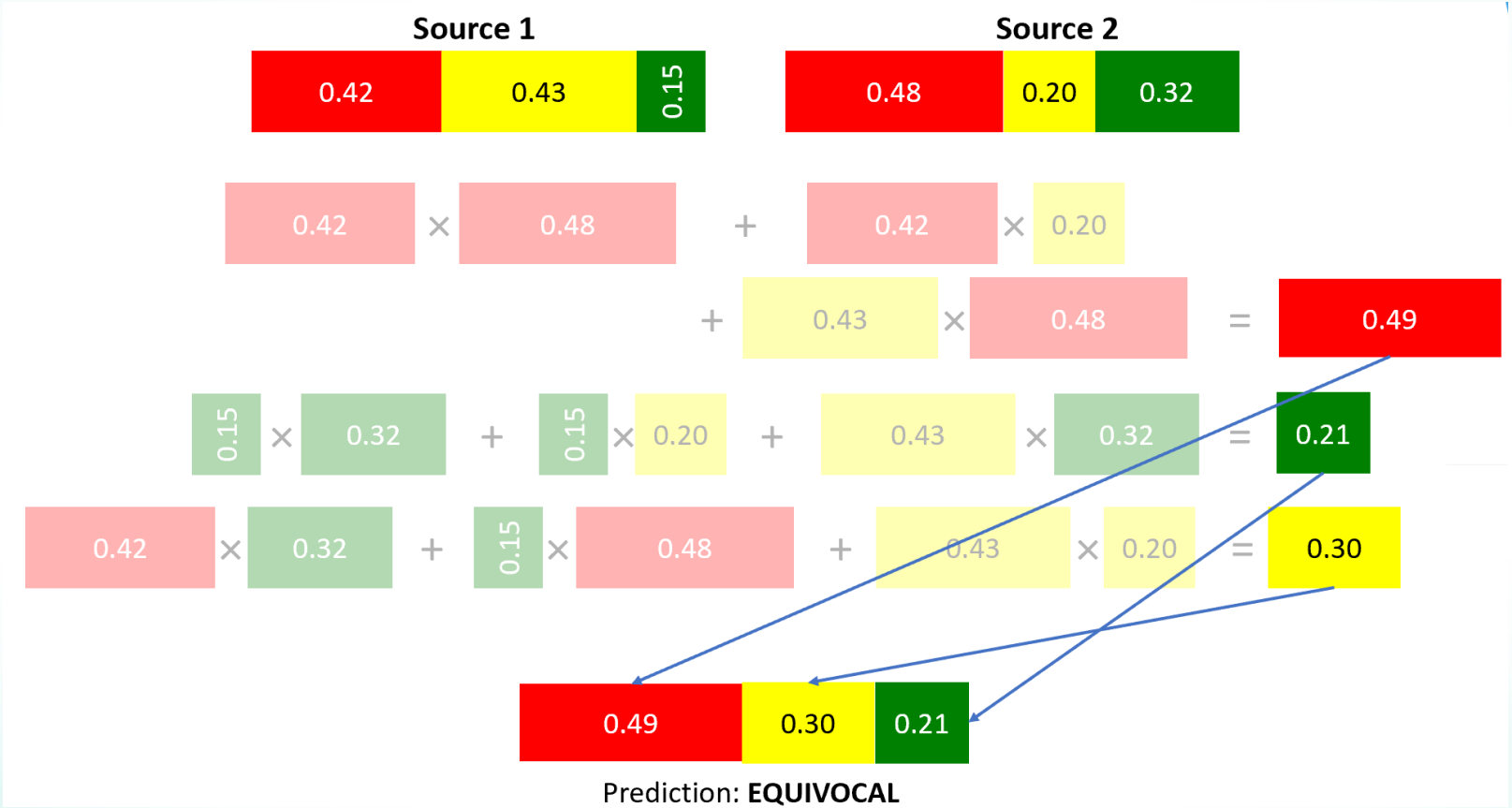
Dempster Combination Rule



Yager Combination Rule

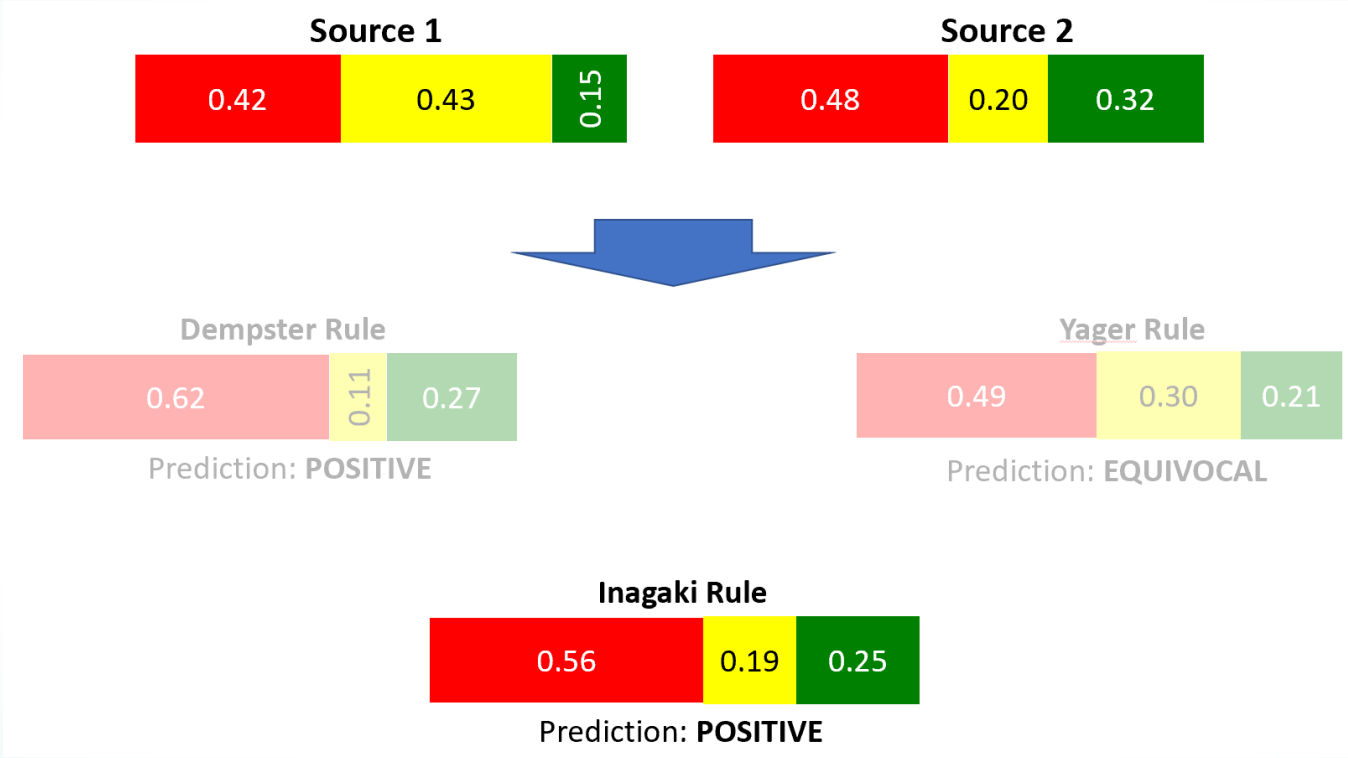


Yager Combination Rule

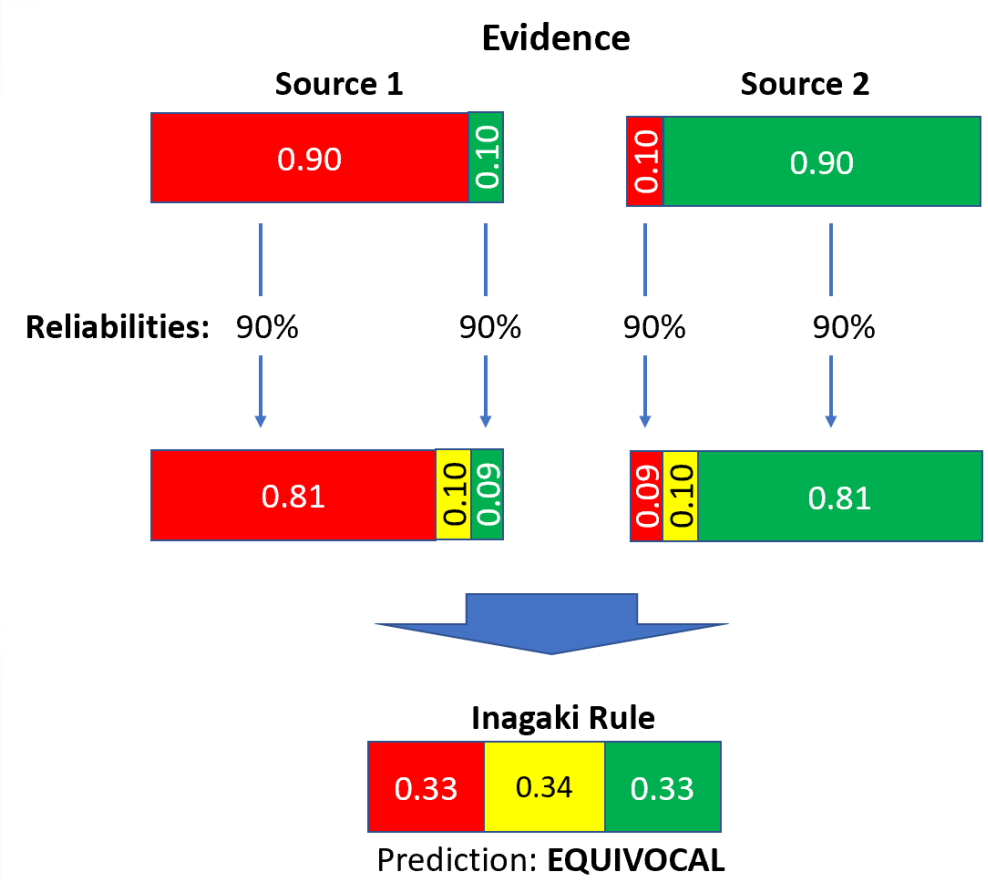


Prediction: **EQUIVOCAL**

Inagaki Combination Rule

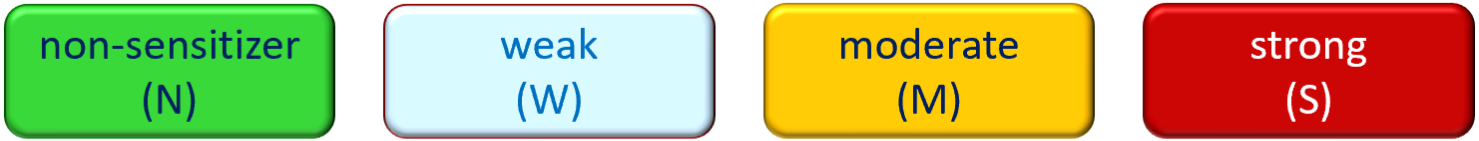


When Evidence Sources Disagree



Ordinal Classification

Consider a four-level classification model for skin sensitization:

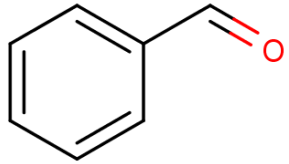


The focal elements can be defined such that the model has 8 possible prediction outcomes:

DST allows us to capture different degrees of uncertainty.

Quantifying Uncertainty of *In Vitro* Assays

Example: skin sensitization (LLNA) for benzaldehyde



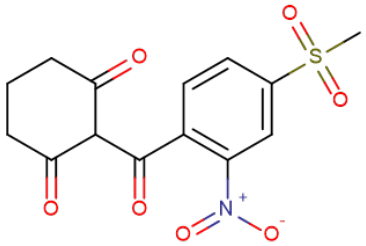
Performance metrics

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

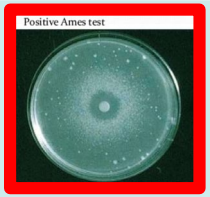
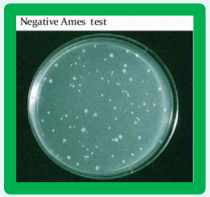
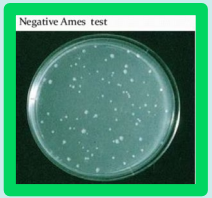
(Urbisch et al., 2015)

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

Combining Evidence from Multiple Ames Assays



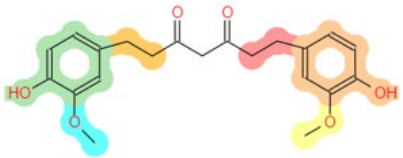
Mesotrione

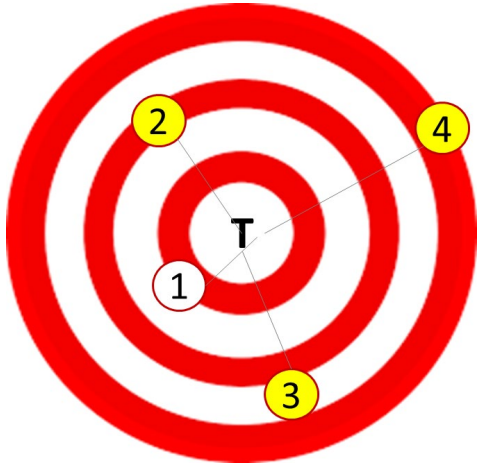
Ames Assay Result	Study Reliability	Probability Bounds	DST Combination
	0.50	0.50	0.74
	0.95	0.95	**WoE Outcome** **NEGATIVE**
	0.80	0.80	

Reliability Measures for Read-Across

- Read-across: using data available for suitable analogs to infer toxicity of a target compound

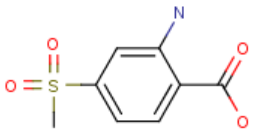
- Reliability depends on (Schultz et al., 2019)
  - Similarity of analog(s) to the target
    - structure similarity
    - property similarity
    - metabolic similarity
    - toxicodynamic similarity
    - toxicokinetic similarity



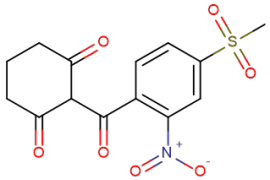


  - reliability of tox data available for analogs

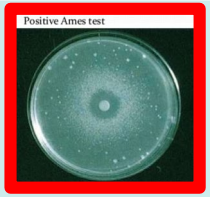
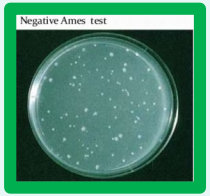
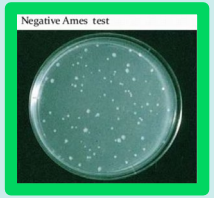
Combining Evidence from an Analog for Read-Across



target:
metabolite of
Mesotrione

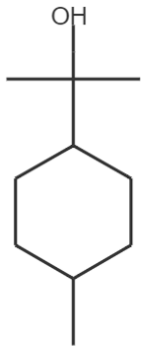
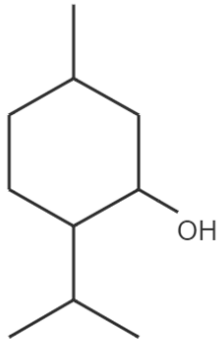
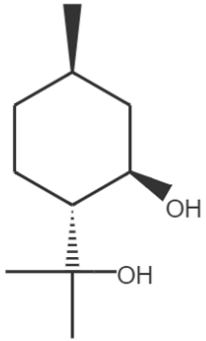


analog:
Mesotrione

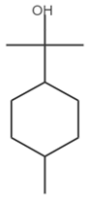
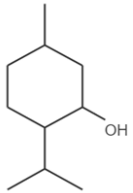
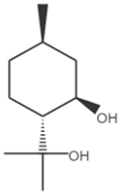
Ames Assay Result	Study Reliability	Analog Quality	Probability Bounds	DST Combination
	0.50	0.62	0.31	0.66
	0.95		0.59	**WoE Outcome** **NEGATIVE**
	0.80		0.50	

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
		

ChemTunes • ToxGPS

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			
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		

Dealing with Uncertainty in the Real World

We are all experienced safety and risk assessors!

Uncertainty About Uncertainty

- We want to make good decisions using methods that are
  - Transparent
  - Interpretable and mechanistic
  - As simple as possible

- Although we often reason probabilistically and understand that certain evidence may be more reliable than other, we may not feel confident in our ability to *quantify* these.

- We are often uncomfortable reporting decisions with any appreciable uncertainty, or if there are conflicting pieces of evidence.

Subjectivity

- When dealing with complex problems, we should expect that experts evaluating the same evidence will not always agree.
- Differences will in part be due to how each person evaluates the reliability and relevance of different sources of evidence.
- Probability bounds analysis provides a systematic and rigorous approach to help focus the conversation and identify *why* we reach different conclusions.

Communicating Uncertainty: Language Matters!

- Explore different ways to communicate
  - pictograms
  - verbal qualifiers: “likely,” “perhaps”
  - numeric ranges: “50 to 70%,” “less than 70%,” “at most 70%”…

- Studies suggest communicating uncertainty using natural frequencies instead of probabilities can be effective: “3 out of 5” or “36 out of 60” instead of 60%.

Communicating uncertainty is an interesting and active area of study in Psychology, Public Policy, and Engineering.

References

- Cronin, M.T.D., Richarz, A., Schultz, T. “Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction,” *Regulatory Toxicology and Pharmacology* 106, 90–104 (2019)
- International Conference on Uncertainty in Risk Analysis: Challenges and Advances in Assessing, Managing and Communicating Uncertainty, Sponsored by Bundesinstitute für Risikobewertung (BfR) and European Food Safety Authority (EFSA), Berlin, Feb 20-22, 2019.
- Klimisch, H., Andreae, M., Tillmann, U. “A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data,” *Regulatory Toxicology and Pharmacology* 25, 1–5 (1997).
- 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).
- Schultz, T., Richarz, A., Cronin, M.T.D. “Assessing uncertainty in read-across: Questions to evaluate toxicity predictions based on knowledge gained from case studies.” *Computational Toxicology*, 9, 1-11 (2019).
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H. “Assessing skin sensitization hazard in mice and men using non-animal test methods,” *Regulatory Toxicology and Pharmacology 71*, 337-351 (2015).

Acknowledgements

MN-AM (Molecular Networks GmbH and Altamira LLC)

The Ohio State University

Ohio State University