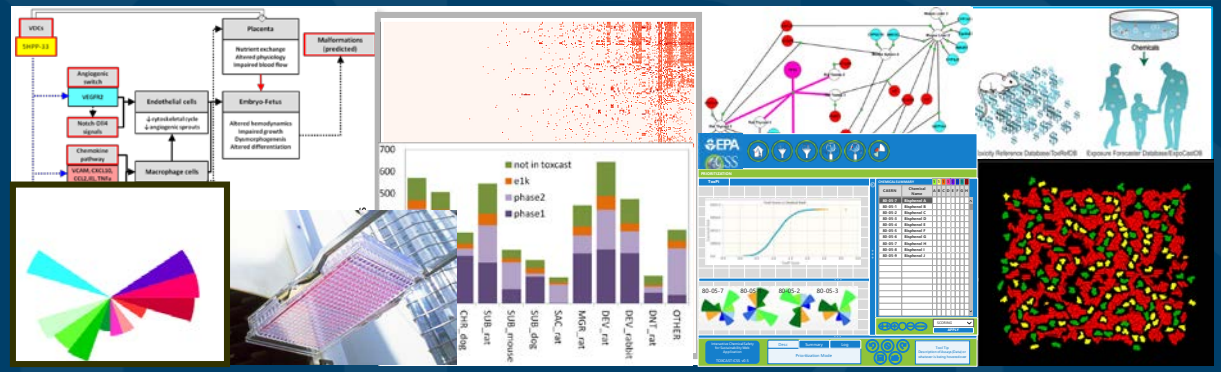
Actualizing research into practical tools: A case study of GenRA, a new read-across tool



George Helman[1,2], Imran Shah[2], Grace Patlewicz[2]

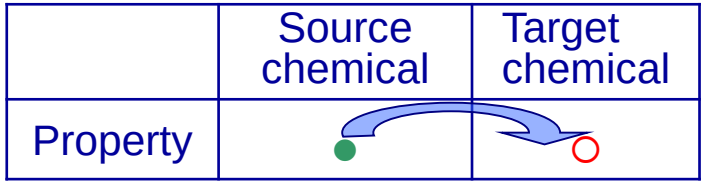
[1]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA
[2]National Center for Computational Toxicology, US EPA, RTP, NC, USA

Outline

- Definitions
- Introduction to the Generalized Read-Across (GenRA) approach
- Current progress on web tool development
- Future work on web tool development
- Conclusions

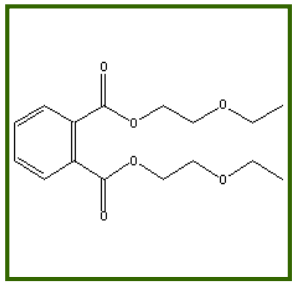
Definitions: Read-across

- Read-across describes the method of filling a data gap whereby a chemical with existing data values is used to make a prediction for a ‘similar’ chemical.
- A target chemical is a chemical which has a data gap that needs to be filled i.e. the subject of the read-across.
- A source analogue is a chemical that has been identified as an appropriate chemical for use in a read-across based on similarity to the target chemical and existence of relevant data.



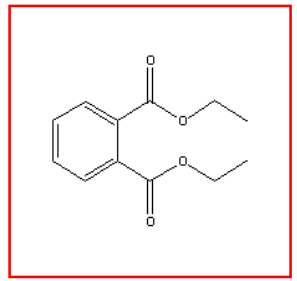
● Reliable data

○ Missing data



Known to be harmful

Acute toxicity?



Predicted to be harmful

GenRA - Introduction

- GenRA (Generalized Read-Across) is a “local validity” approach predicting toxicity as a similarity-weighted activity of source analogues based on chemistry and/or bioactivity descriptors. (Shah et al, 2016)
- Generalized version of Chemical-Biological Read-Across (CBRA) developed by Low et al (2013)

- **Goal**: to establish an objective performance baseline for read-across and quantify the uncertainty in the predictions made.

- Tested and compared chemical, bioactivity (hitcalls from ToxCast) descriptors as well as a hybrid of the two.
- Underlying data used was taken from ToxRefDB, a collection of repeated dose toxicity study types e.g. chronic, multigeneration, developmental, subchronic etc
- Toxicity effects within those study types were recorded as binary outcomes

Algorithm

1. Select chemical/biological fingerprint to characterize target and source analogues. The Jaccard similarity metric is used to quantify the pairwise similarity
2. Choose the number of source analogues, k (where k = the number of nearest neighbors)
3. Filter on the basis of *in vivo* toxicity data for source analogues
4. Read-across prediction for target chemical is the outcome of a similarity-weighted activity calculation from the nearest neighbors
5. Perform Receiver Operating Characteristic (ROC) to identify the threshold where the best performance is produced for a given s and k.

Jaccard similarity:

$$s_{ij} = \frac{\Sigma_l (x_{il} \wedge x_{jl})}{\Sigma_l (x_{il} \vee x_{jl})}$$

$$y_i = \frac{\sum_j^k s_{ij} x_j}{\sum_j^k s_{ij}}$$

Results of GenRA

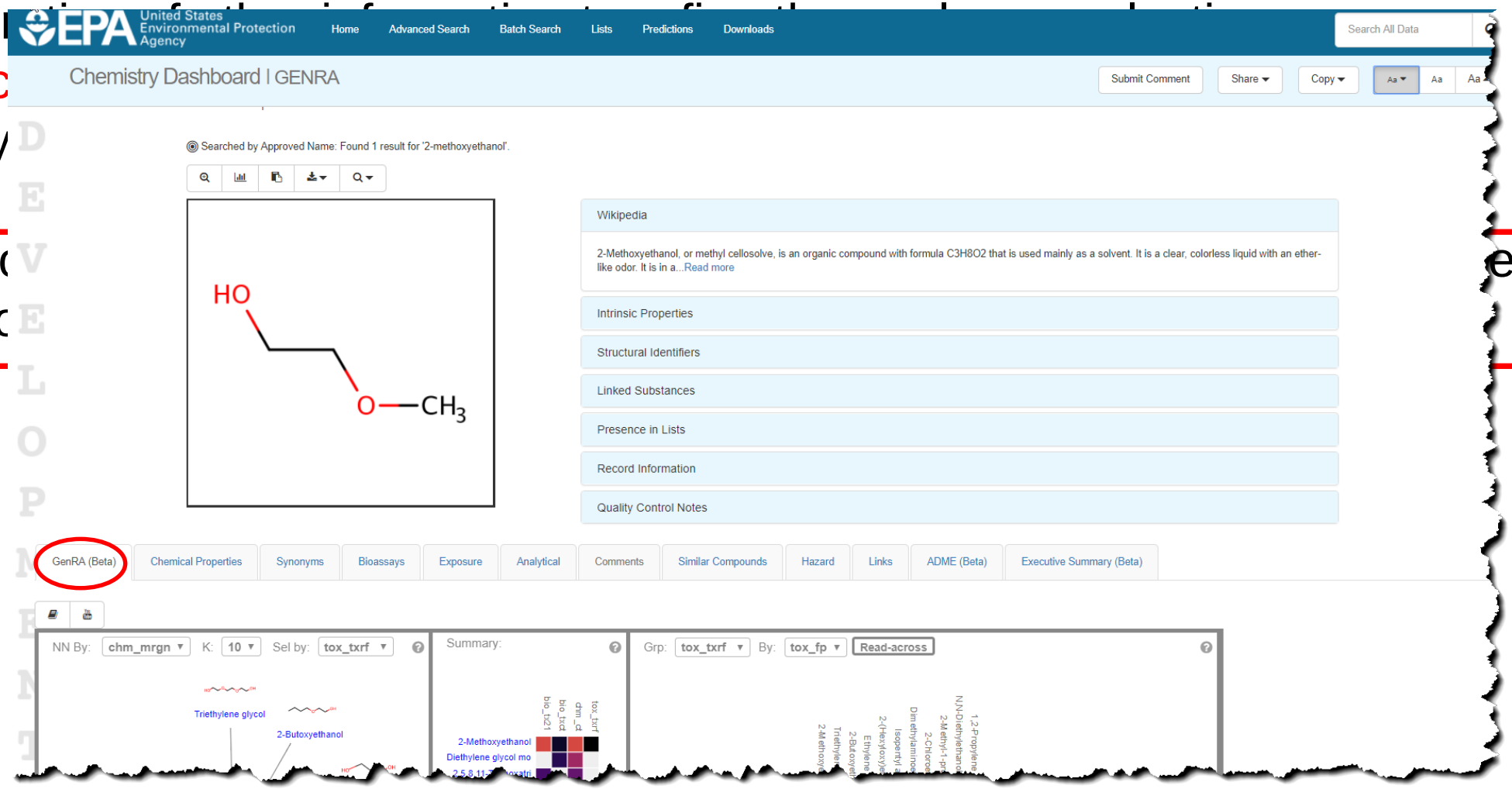
Tabulation of high-performing scores ($p<0.01$)

Study	Cases*	chm	bio	bc	chm > bio| bc	bio > chm|bc	bc > chm|bio	Any
chr	1115	97 (9%)	133 (12%)	67 (6%)	75 (7%)	107 (10%)	21 (2%)	203 (18%)
dev	712	63 (9%)	138 (19%)	50 (7%)	43 (6%)	117 (16%)	7 (1%)	167 (23%)
dnt	84	16 (19%)	26 (31%)	15 (18%)	5 (6%)	20 (24%)	2 (2%)	27 (32%)
mgr	668	78 (12%)	126 (19%)	56 (8%)	46 (7%)	106 (16%)	4 (1%)	156 (23%)
rep	34	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
sac	128	2 (2%)	7 (5%)	6 (5%)	1 (1%)	6 (5%)	2 (2%)	9 (7%)
sub	922	107 (12%)	140 (15%)	78 (8%)	64 (7%)	122 (13%)	21 (2%)	212 (23%)
All	3663	363 (10%)	570 (16%)	272 (7%)	234 (6%)	478 (13%)	57 (2%)	774 (21%)

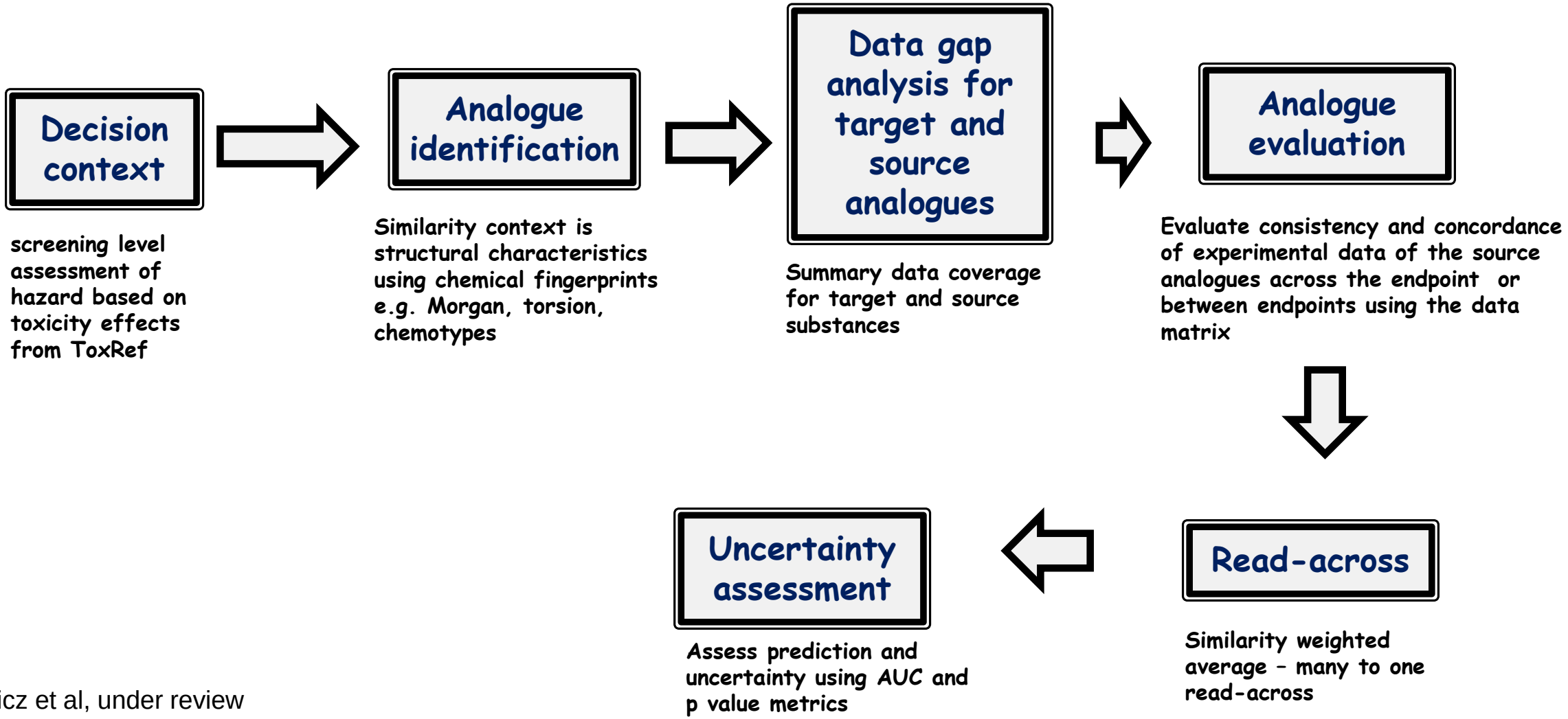
*cluster-study-effect combination

Next steps - GenRA

- Ongoing analysis:
- Consider
  physicoc
  similarity
- Potency
- Integrati
  https://cc

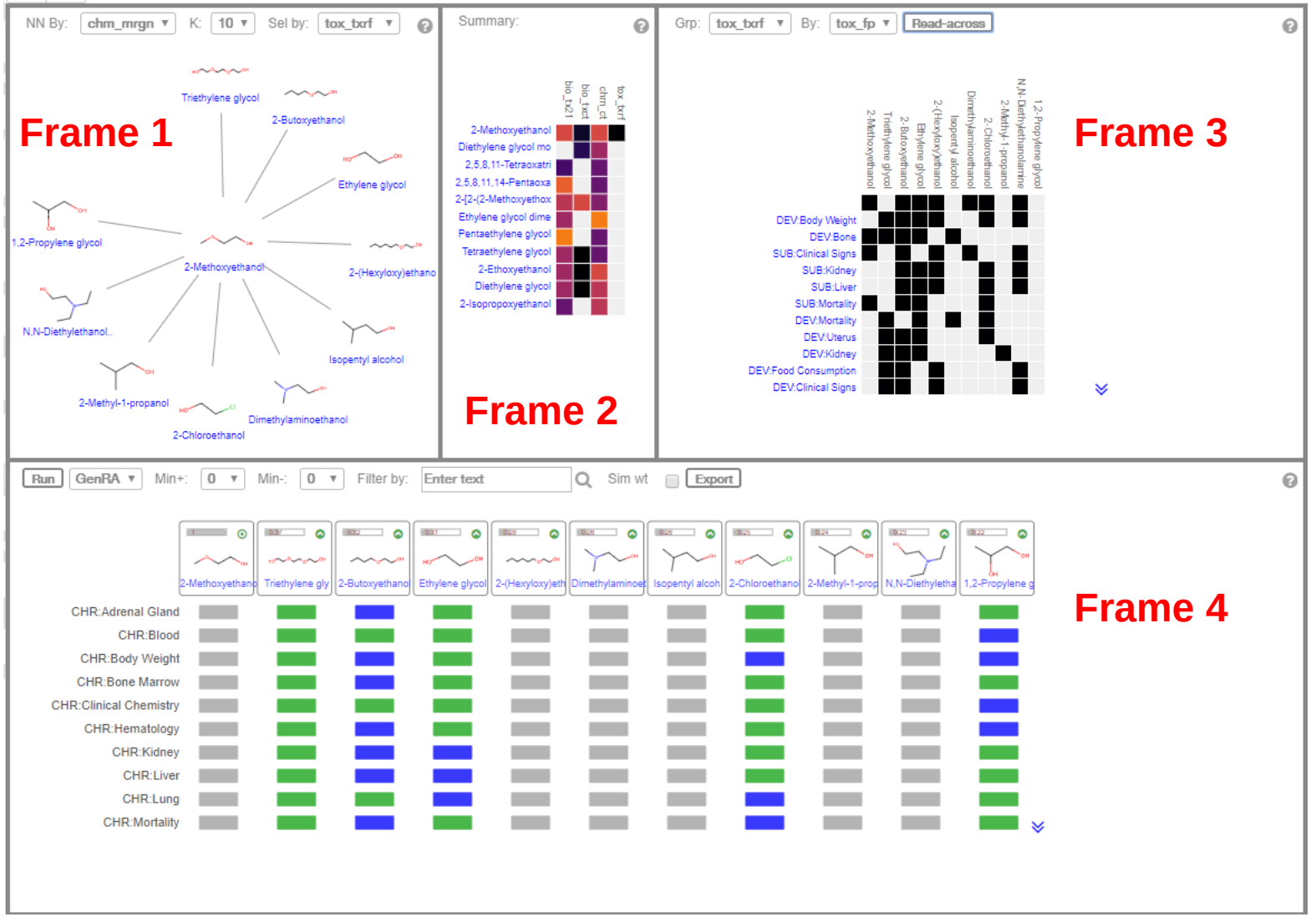


Category Workflow and its alignment to GenRA

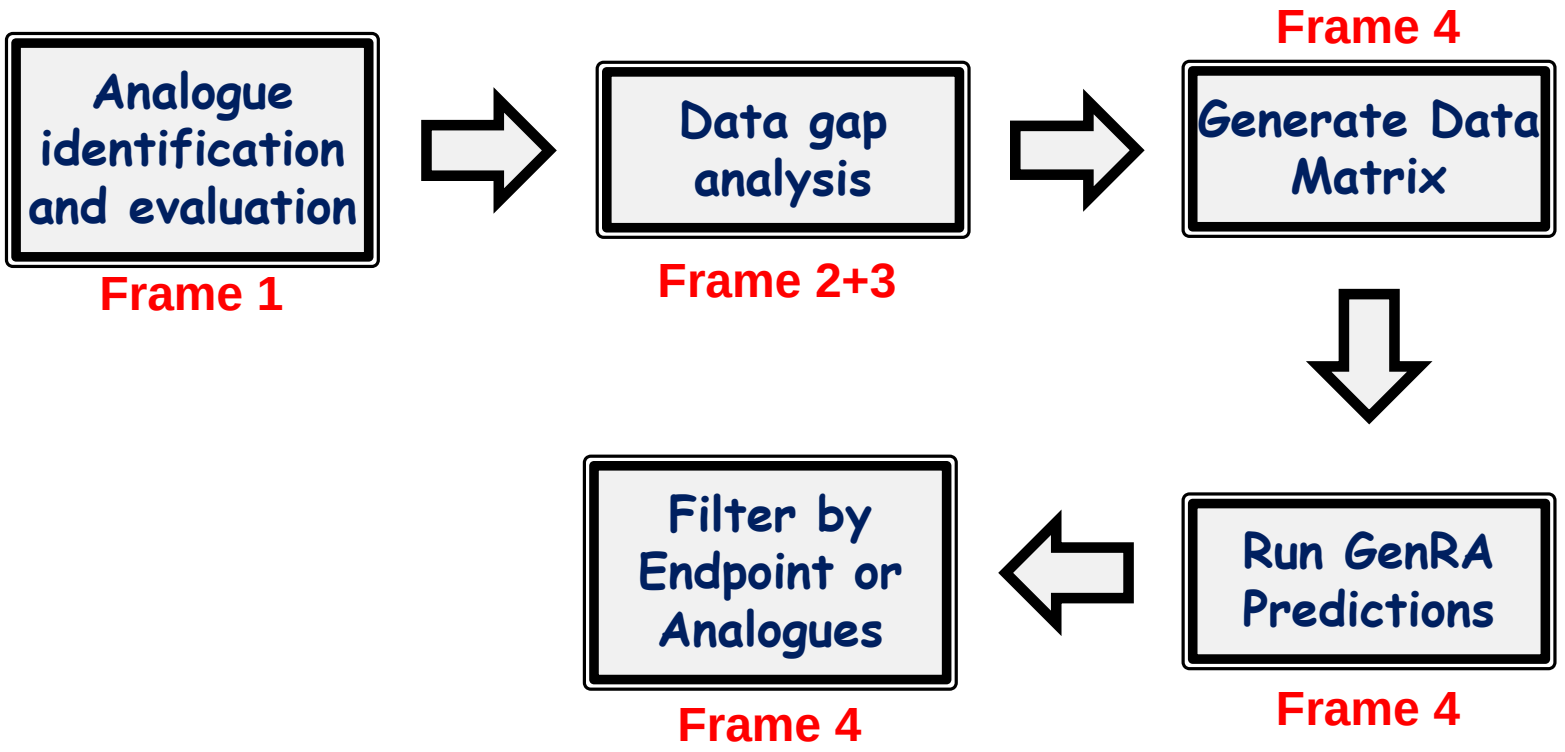


Patlewicz et al, under review
Helman et al, in prep

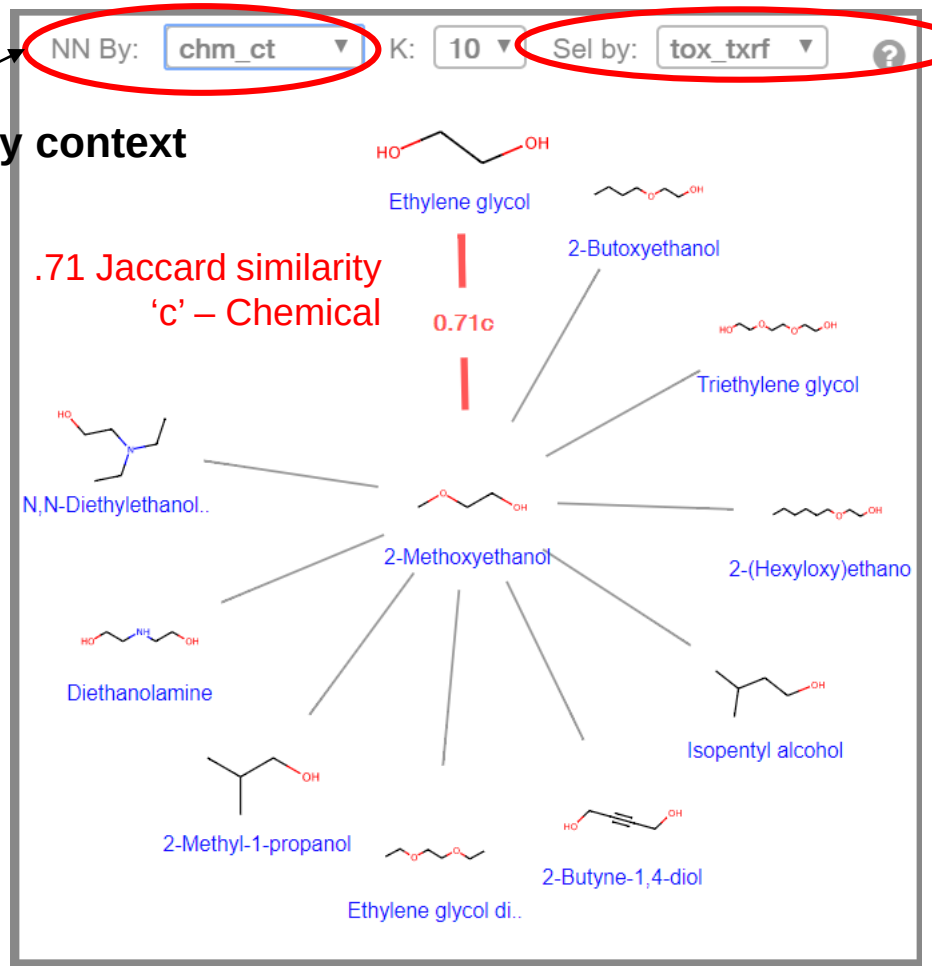
GenRA: Web Tool Development



Web Tool Workflow

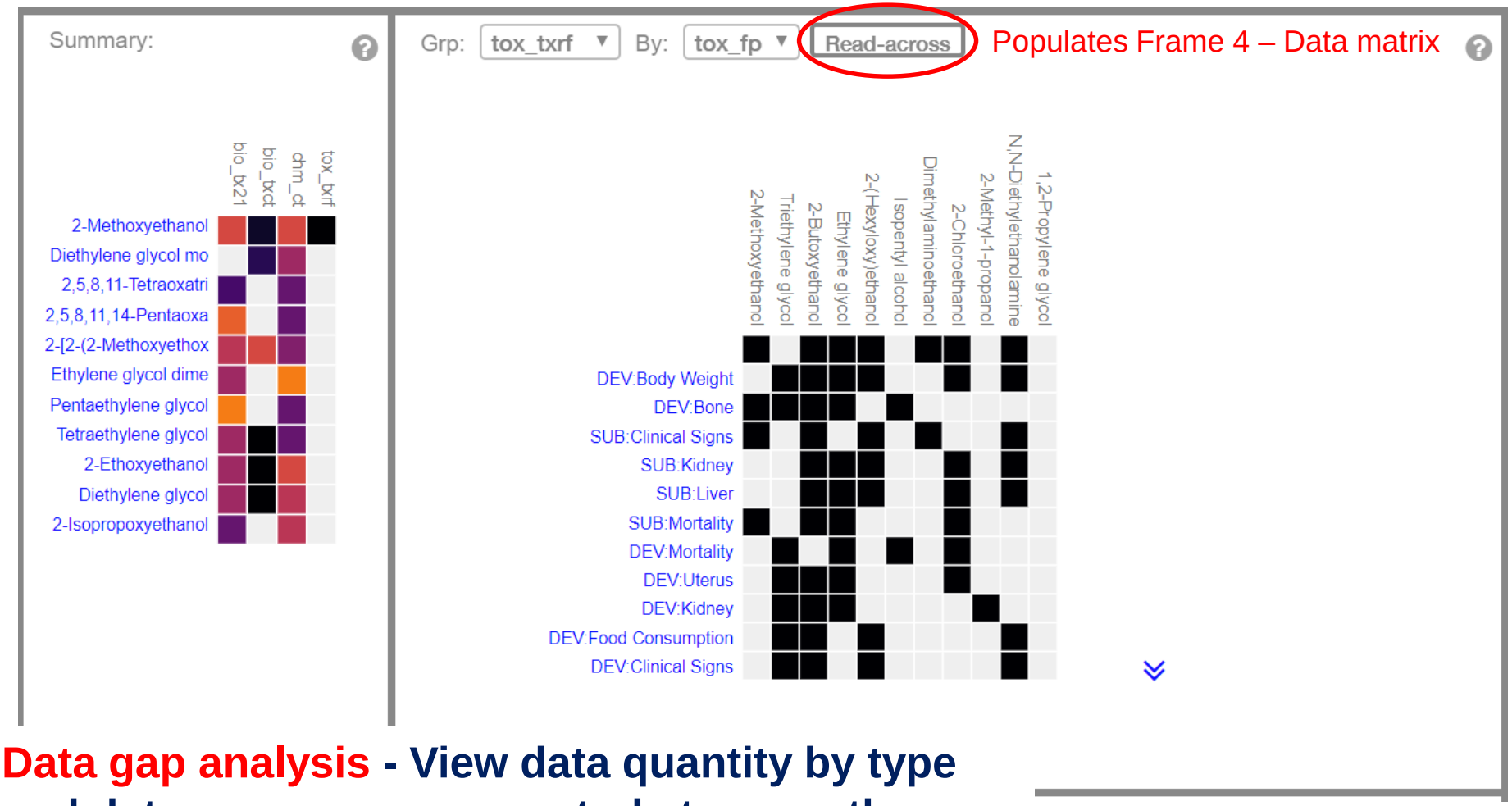


Analogue Identification and Evaluation



- 'NN By' option determines fingerprint of choice i.e. search for source analogues on the basis of different chemical fingerprints.
- 'K' option determines the number of source analogues to search for
- 'Sel by' option subsets chemical space to search for source analogues that have certain types of data, by default *in vivo* data.

Data Gap Analysis



- 'Grp' option controls the data that is being shown in the heatmap.
- 'By' option allows you to pivot the heatmap to look at different views of the same data.

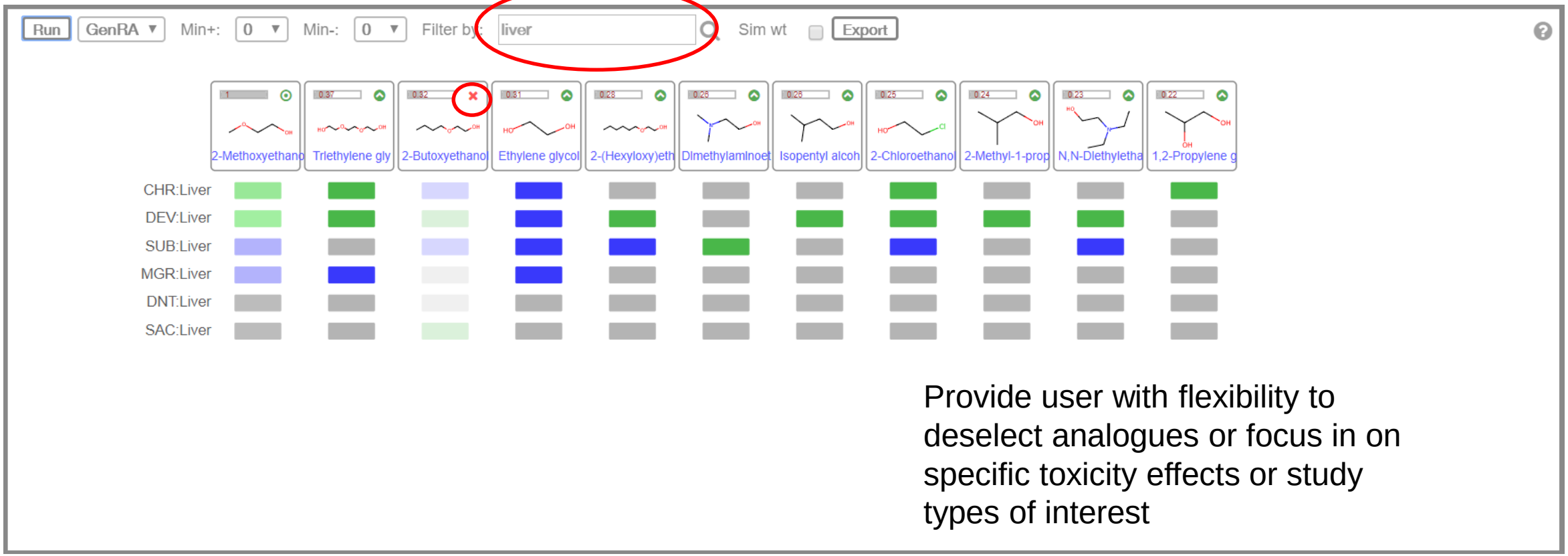
Data gap analysis - View data quantity by type and data coverage across study type on the basis of toxicity effects

Generate Data Matrix

- 'Min+' and 'Min-' options control minimum number of pos and neg values in displayed endpoints.
- 'Sim wt' option changes width of cells to scale of the similarity of the neighbor.
- 'Export' will export data matrix to CSV

Run GenRA Predictions

Filter by Endpoint or Analogues



Provide user with flexibility to deselect analogues or focus in on specific toxicity effects or study types of interest

Future Work: Physicochemical Similarity Context

- Searching on the basis of physicochemical information in addition to structure can improve GenRA performance (Helman et al, 2018)
- Physicochemical information will be incorporated into the analogue identification and evaluation steps of the workflow.
- The ‘Lipinski Rule of 5’ properties (LogP, MW, #HB donors, #HB acceptors) were used in the analysis undertaken.

Future Work: Point of Departure Prediction

- Algorithm readily extendible to continuous data.

$$y_i = \frac{\sum_j^k s_{ij} x_j}{\sum_j^k s_{ij}}$$

- Preliminary results appear promising for predicting acute oral rat LD50 data.

Conclusions

- GenRA is an objective algorithmic approach for generating read-across predictions.
- A web tool has been developed to front-end the GenRA workflow
- The web tool will be made available on NCCT's CompTox dashboard soon.

Acknowledgements

- Imran Shah
- Grace Patlewicz
- Tony Williams
- Jeff Edwards
- Chris Grulke