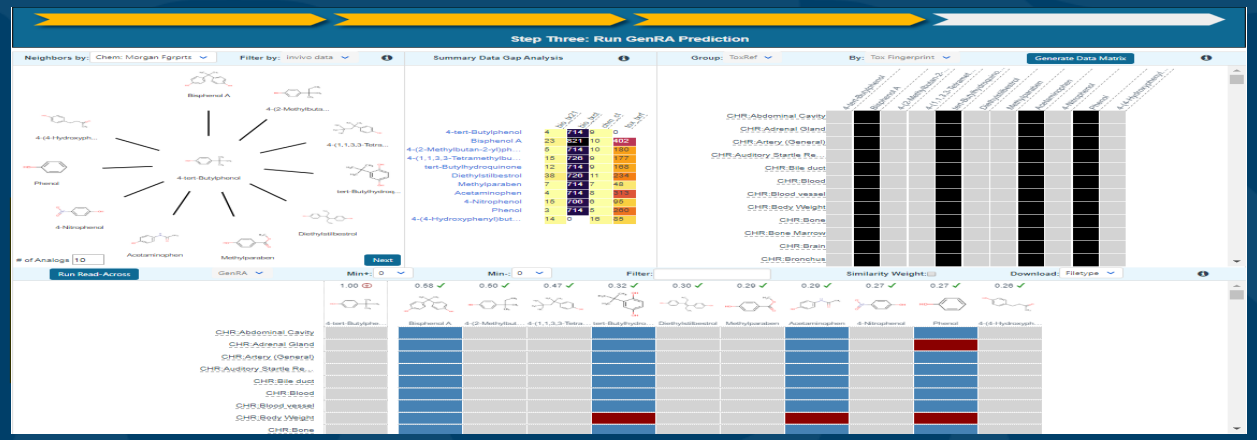
Use of GenRA to assist in read-across



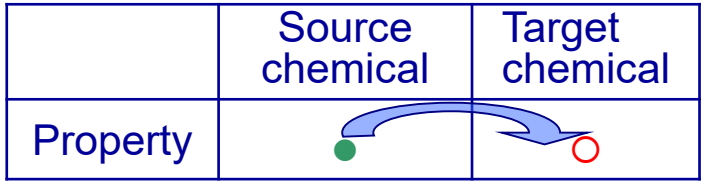
Grace Patlewicz
National Center for Computational Toxicology (NCCT), US EPA

Outline

- What is Read-across?
- What tools exist that might help facilitate read-across?
- Putting those tools into context of the read-across workflow
- Evolving the read-across workflow to address other New Approach Methods (NAMs)
- Generalised Read-across (GenRA) approach
- GenRA implementation
- From theory to practical application
- Ongoing research to enhance GenRA
- Summary remarks

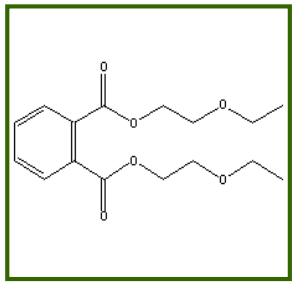
Definitions: Read-across

- 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.



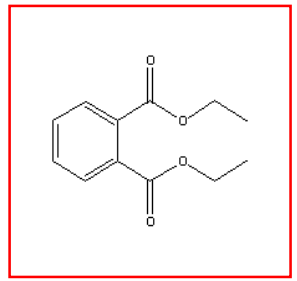
● Reliable data

○ Missing data



Known to be harmful

Acute toxicity?



Predicted to be harmful

Selected publicly read-across tools

Computational Toxicology 3 (2017) 1–18

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

Navigating through the minefield of read-across tools: A review of in silico tools for grouping

Grace Patlewicz [a,*], George Helman [a,b], Prachi Pradeep [a,b], Imran Shah [a]

[a] National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency, 109 TW Alexander Dr, Research Triangle Park (RTP), NC 27711, USA
[b] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 22 May 2017
Accepted 25 May 2017
Available online 29 May 2017

Keywords:
Category approach
Analogue approach
Data gap filling
Read-across
(Q)SAR
Trend analysis
Nearest neighb

ABSTRACT

Read-across is a popular data gap filling technique used within analogue and category approaches for regulatory purposes. In recent years there have been many efforts focused on the challenges involved in read-across development, its scientific justification and documentation. Tools have also been developed to facilitate read-across development and application. Here, we describe a number of publicly available read-across tools in the context of the category/analogue workflow and review their respective capabilities, strengths and weaknesses. No single tool addresses all aspects of the workflow. We highlight how the different tools complement each other and some of the opportunities for their further development to address the continued evolution of read-across.

Published by Elsevier B.V.

Selected publicly read-across tools

Tool	AIM	ToxMatch	AMBIT	OECD Toolbox	CBRA	ToxRead	GenRA
Analogue identification	X	X	X	X	X	X	X
Analogue Evaluation	NA	X	X by other tools available	X	X	X For Ames & BCF	NA
Data gap analysis	NA	X	X Data matrix can be exported	X Data matrix viewable	NA	NA	X Data matrix can be exported
Data gap filling	NA	X	User driven	X	X	X	X
Uncertainty assessment	NA	NA	NA	X	NA	NA	X
Availability	Free	Free	Free	Free	Free	Free	Free

Read-across workflows

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com

Journal Cover Image

Navigating through the minefield of read-across frameworks: A commentary perspective

Grace Patlewicz[a, *], Mark T.D. Cronin[b], George Helman[a, c], Jason C. Lambert[d], Lucina E. Lizarraga[d], Imran Shah[a]

[a] National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency (US EPA), 109 TW Alexander Dr, Research Triangle Park (RTP), NC 27711, USA
[b] School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK
[c] Oak Ridge Institute for Science and Education (ORISE), 1299 Bethel Valley Road, Oak Ridge, TN 37830, USA
[d] National Center for Evaluation Assessment (NCEA), US Environmental Protection Agency (US EPA), 26 West Martin Luther King Dr, Cincinnati, OH 45268, USA

A harmonised hybrid read-across workflow

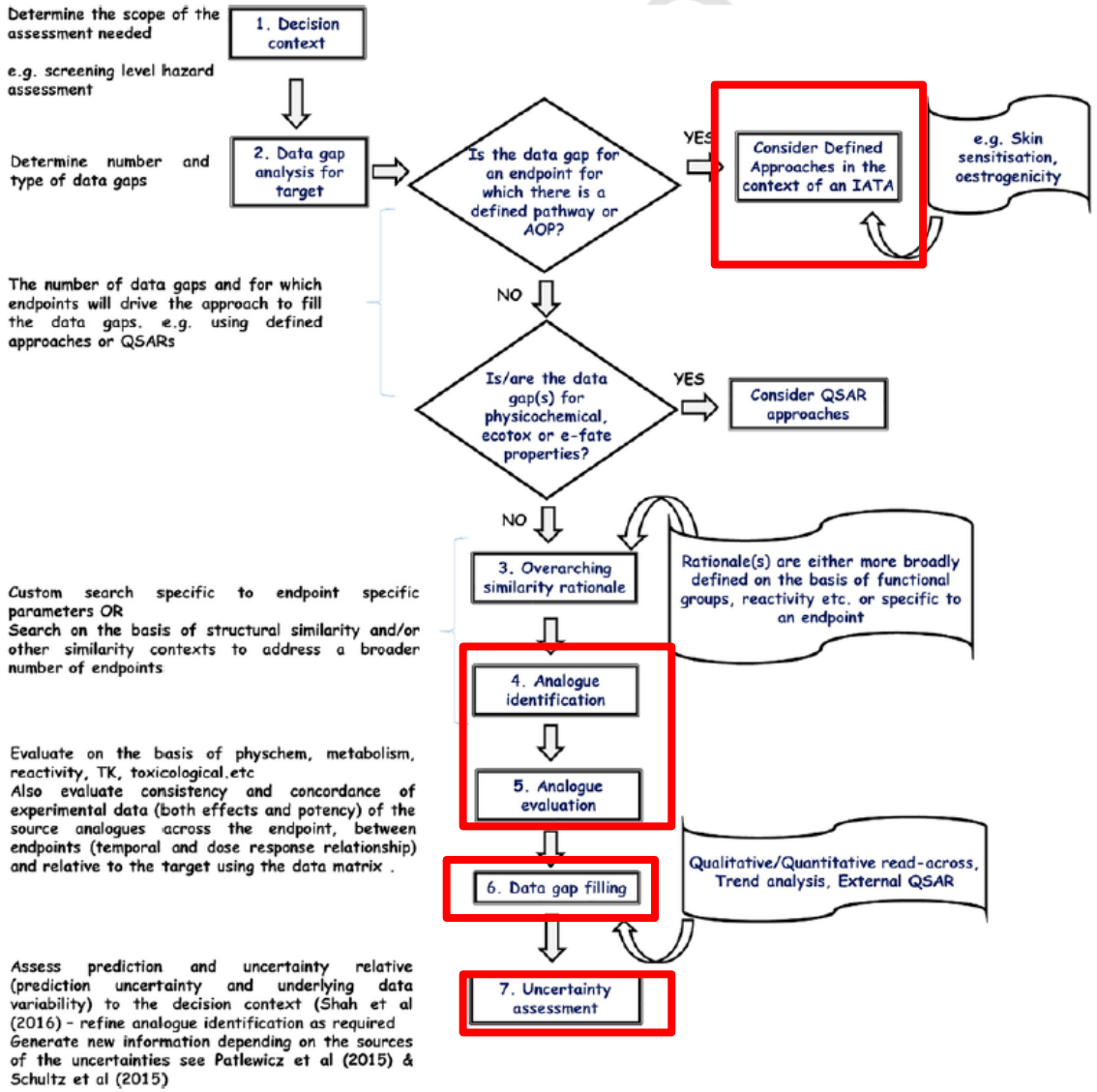


Fig. 9. A harmonised hybrid development and assessment framework.

Where do other NAMs fit?
Current read-across approaches are expert driven?
How should we transition to data-driven approaches?
What about characterising the uncertainty of the predictions made?

Patlewicz et al., 2018

Generalised Read-Across: GenRA

- Predicting toxicity as a similarity-weighted activity of nearest neighbours based on chemistry and/or bioactivity descriptors
- Systematically evaluates read-across performance and uncertainty using available data

Jaccard similarity:

GenRA v1 - Approach

I. Data

1,778 Chemicals
3,239 Structure descriptors (chm)
820 Bioactivity hitcall (bio) ToxCast

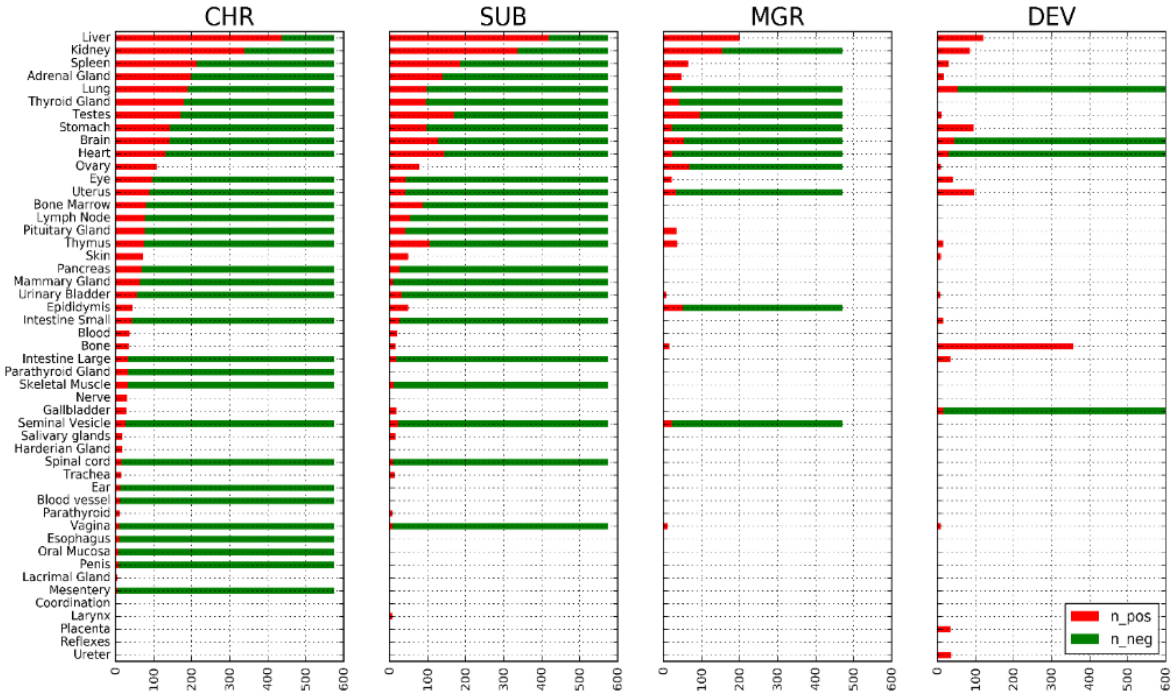
574 toxicity effects (tox) ToxRefDB

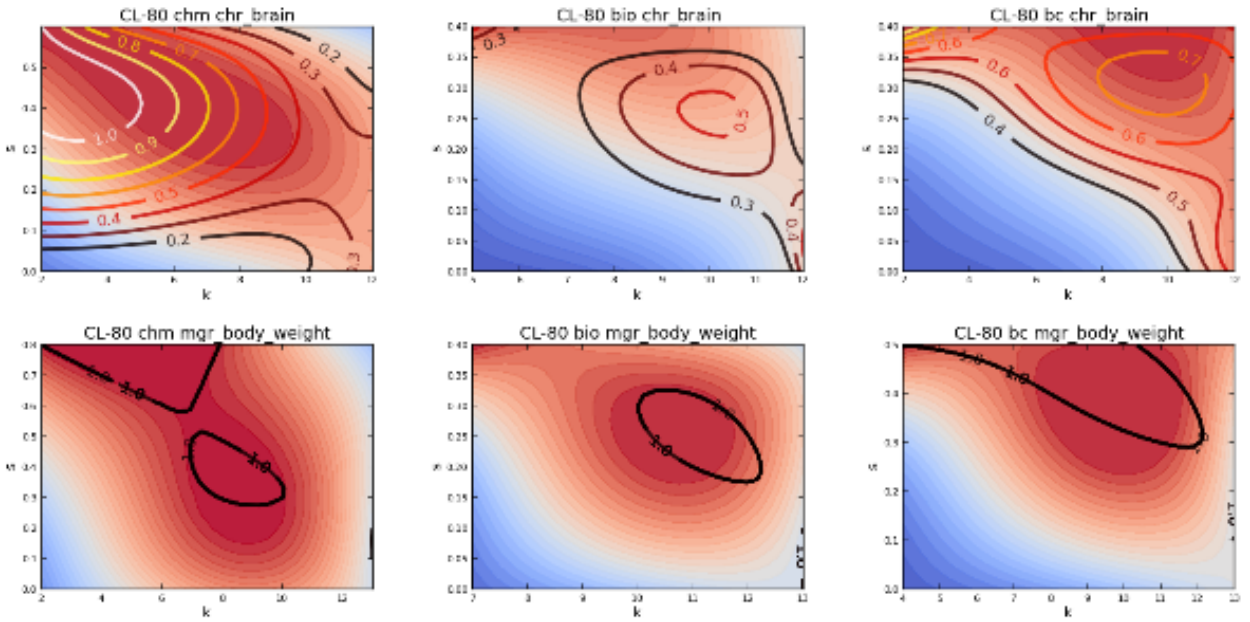
II. Define Local neighbourhoods

Use K-means analysis to group chemicals by similarity
Use cluster stability analysis
~ 100 local neighbourhoods

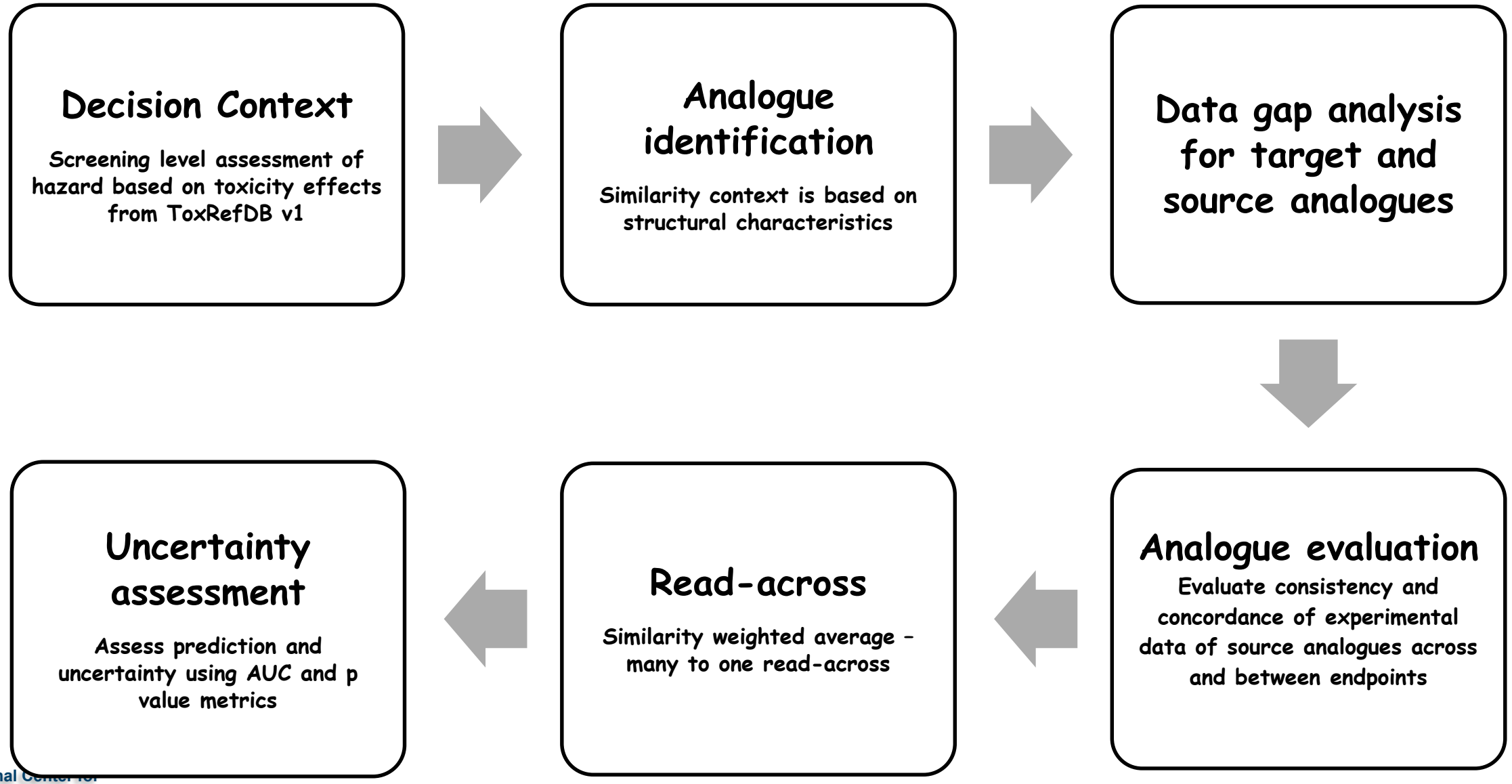
III. GenRA

Use GenRA to predict toxicity effects in local neighbourhoods
Evaluate impact of structural and/or bioactivity descriptors on prediction
Quantify uncertainty





Implementing GenRA within the workflow



GenRA tool in practice

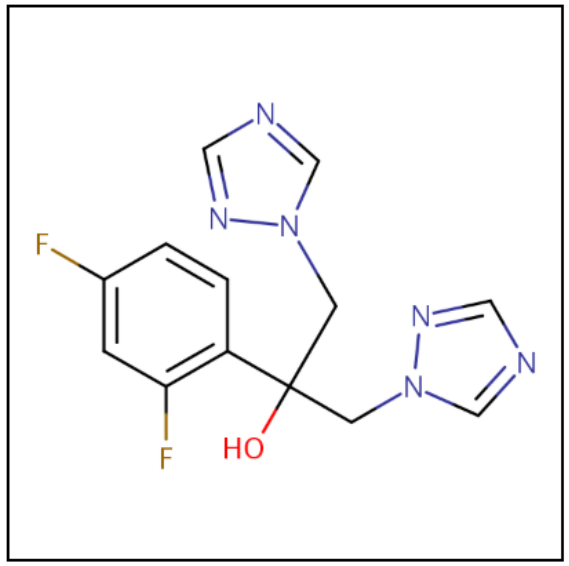
- **Integrated in the EPA CompTox Chemicals Dashboard**

Fluconazole

86386-73-4 | DTXSID3020627

Searched by DSSTox Substance Id.

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS



Wikipedia

Fluconazole is an antifungal medication used for a number of fungal infections. This includes candidiasis, blastomycosis, coccidiodomycosis, cryptococcosis, histoplasmosis, dermatophytosis, and pityriasis versicolor. It is also used to prevent candidiasis in those who are at high risk such as following organ transplantation, low birth weight babies, and those with low blood neutrophil counts. It is given either by mouth or by injection into a vein.

Common side effects include vomiting

...

Read more

Intrinsic Properties

Molecular Formula: $C_{13}H_{12}F_2N_8O$ Mol File

Find All Chemicals

Average Mass: 306.277 g/mol Isotope Mass Distribution

Monoisotopic Mass: 306.104065 g/mol

Structural Identifiers

Linked Substances

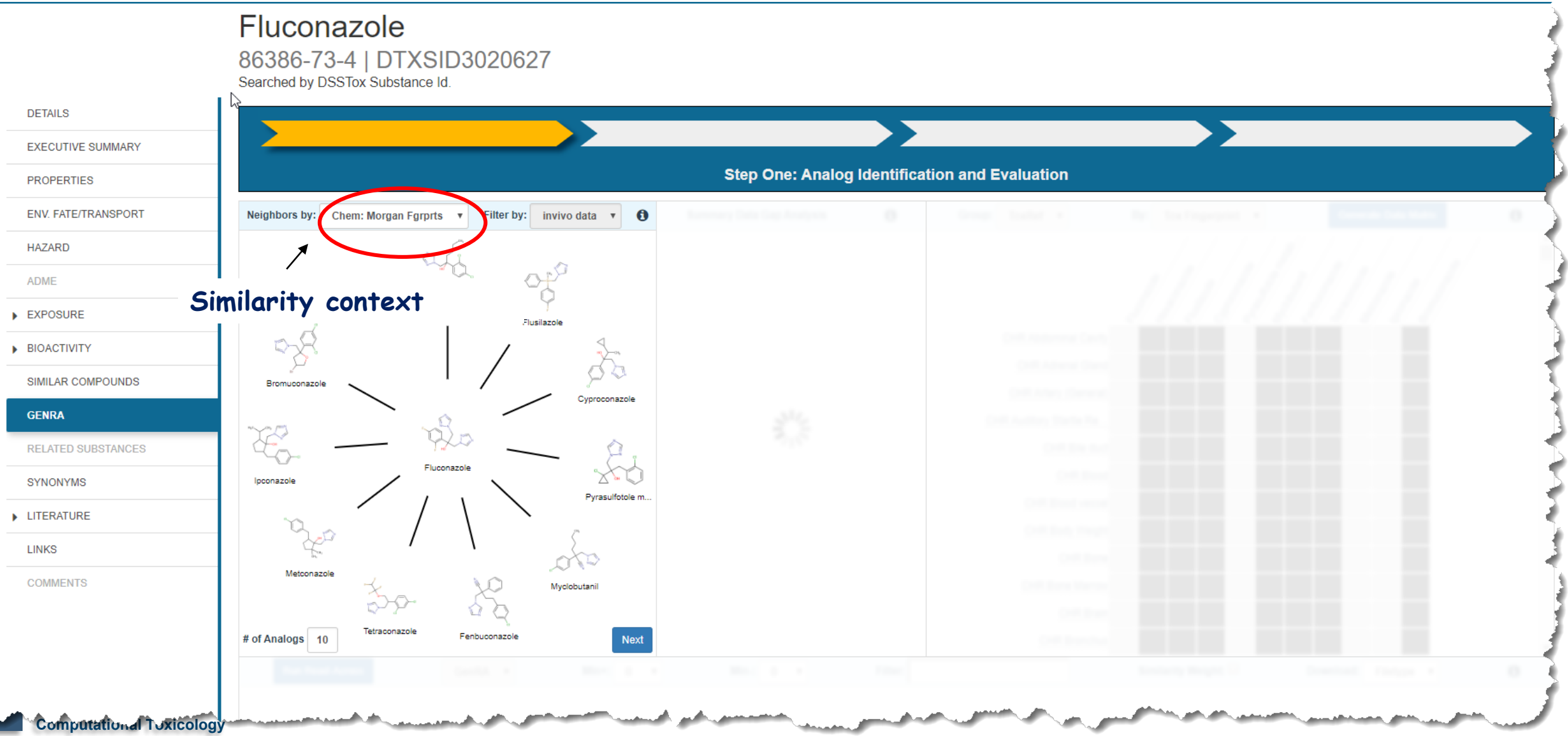
Presence in Lists

Record Information

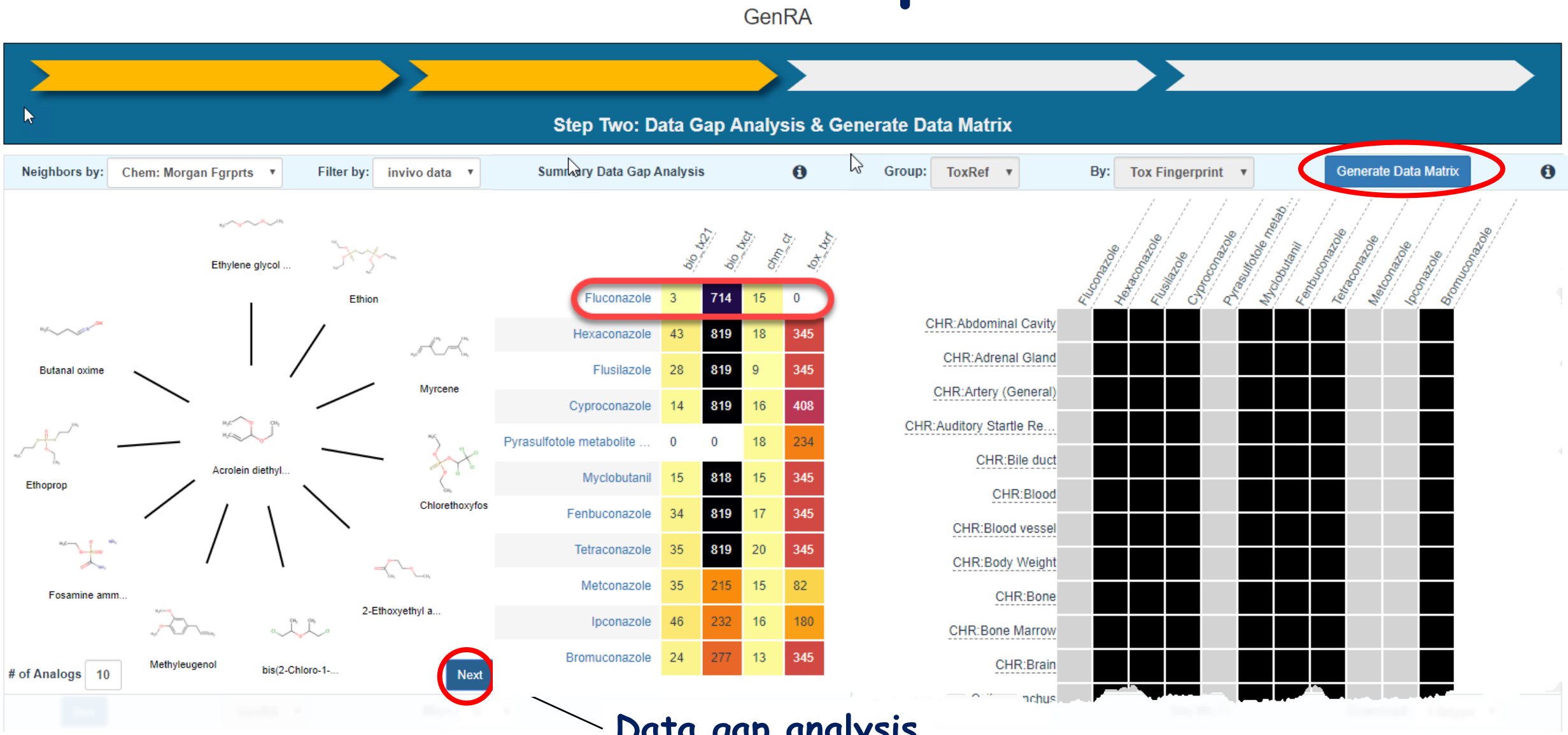
Quality Control Notes

GenRA tool in practice

- Structured as a workflow

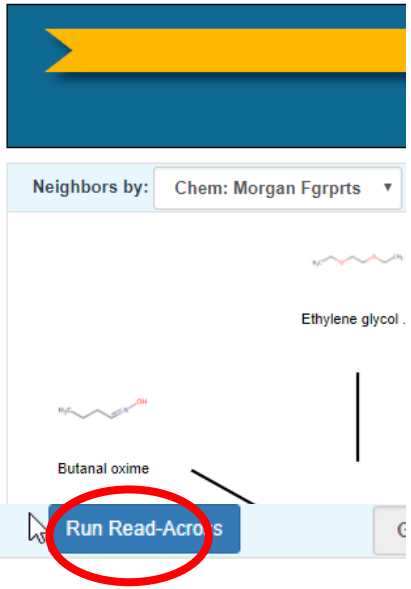


GenRA tool in practice



Data gap analysis

GenRA tool in practice



Run GenRA

ALTEX preprint
published February 4, 2019
doi:10.14573/altex.1811292

Short Communication

Generalized Read-Across (GenRA): A workflow implemented into the EPA CompTox Chemicals Dashboard

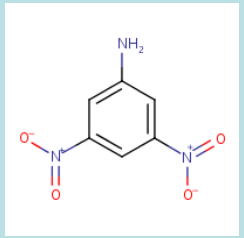
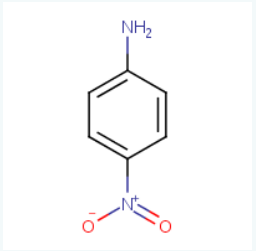
George Helman[1,2], Imran Shah[2], Antony J. Williams[2], Jeff Edwards[2], Jeremy Dunne[2] and Grace Patlewicz[2]*

[1]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA; [2]National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency, Research Triangle Park (RTP), NC, USA

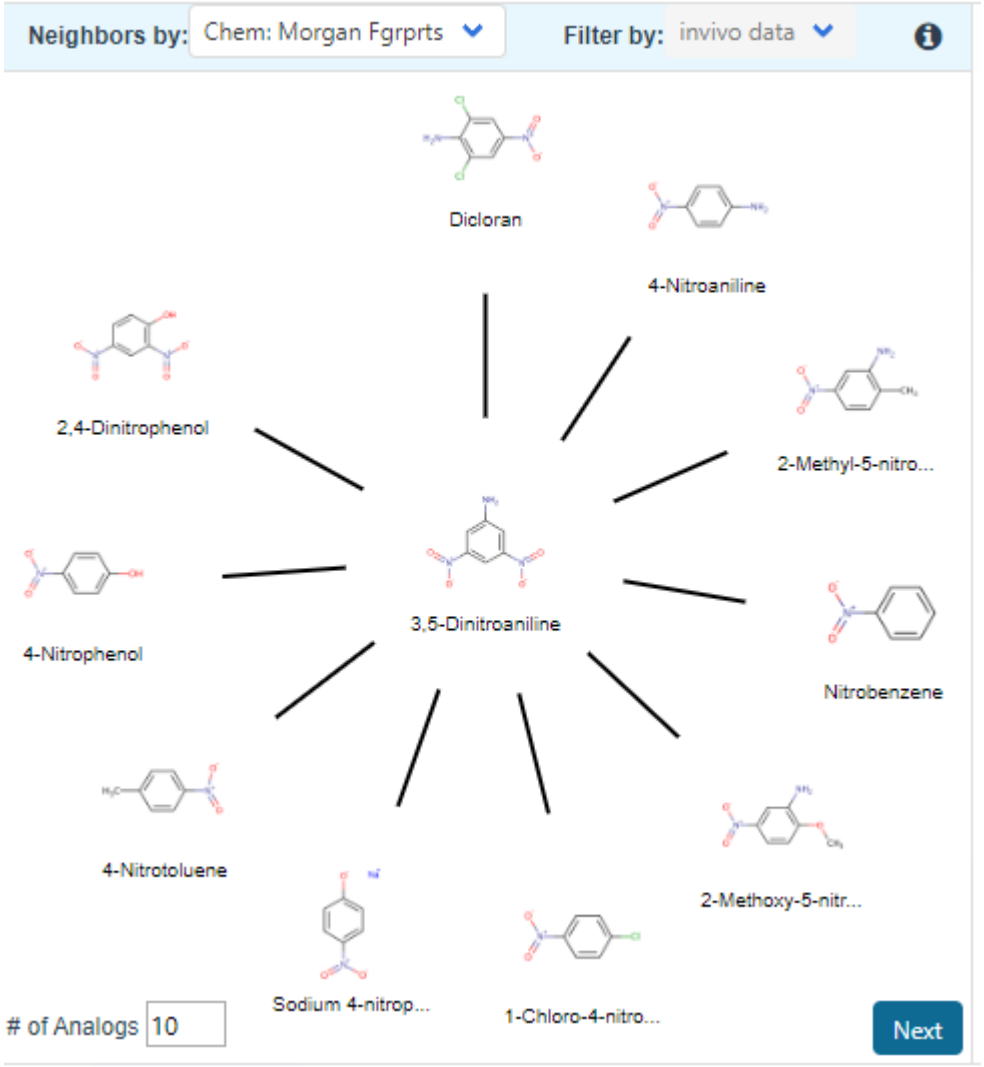
Abstract

Generalized Read-Across (GenRA) is a data driven approach which makes read-across predictions on the basis of a similarity weighted activity of source analogues (nearest neighbors). GenRA has been described in more detail in the literature (Shah et al., 2016; Helman et al., 2018). Here we present its implementation within the EPA's CompTox Chemicals Dashboard to provide public access to a GenRA module structured as a read-across workflow. GenRA assists researchers in identifying source analogues, evaluating their validity and making predictions of *in vivo* toxicity effects for a target substance. Predictions are presented as binary outcomes reflecting presence or absence of toxicity together with quantitative measures of uncertainty. The approach allows users to identify analogues in different ways, quickly assess the availability of relevant *in vivo* data for those analogues and visualize these in a data matrix to evaluate the consistency and concordance of the available experimental data for those analogues before making a GenRA prediction. Predictions can be exported into a tab-separated value (TSV) or Excel file for additional review and analysis (e.g., doses of analogues associated with production of toxic effects). GenRA offers a new capability of making reproducible read-across predictions in an easy-to use-interface.

GenRA in practice – step by step

Target chemical	Proposed source analogue	Primary similarity rationale
Structural		
3,5-Dinitroaniline 	4-Nitroaniline 	Considerations for chemical class, structural moiety, reactivity, metabolism and toxicity were used to refine the pool of analogues. Selection of the source analogue is based on availability of toxicity values, duration of the principal study and health protectiveness of the adopted POD, given the commonalities in the toxicokinetic and toxicity profile for all the candidates.

GenRA v1 in practice – step by step



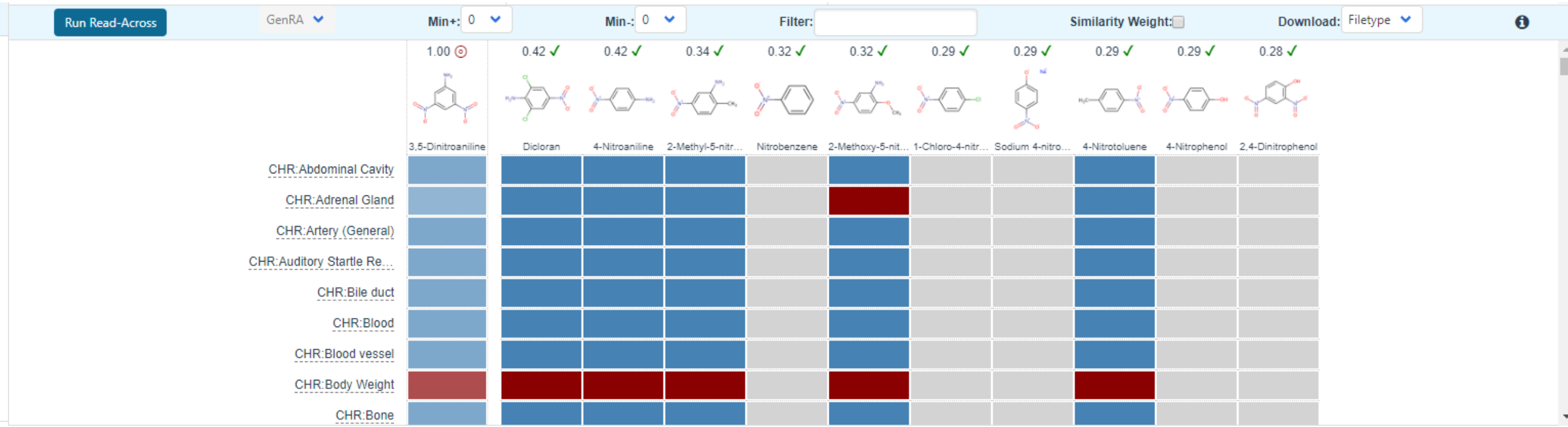
Analogue identification:

Similarity based on Morgan chemical fingerprints and selecting a default of 10 source analogues

GenRA v1 in practice – step by step

Data matrix view of source analogues relative to target chemical

Updated Data matrix view with GenRA predictions for target chemical



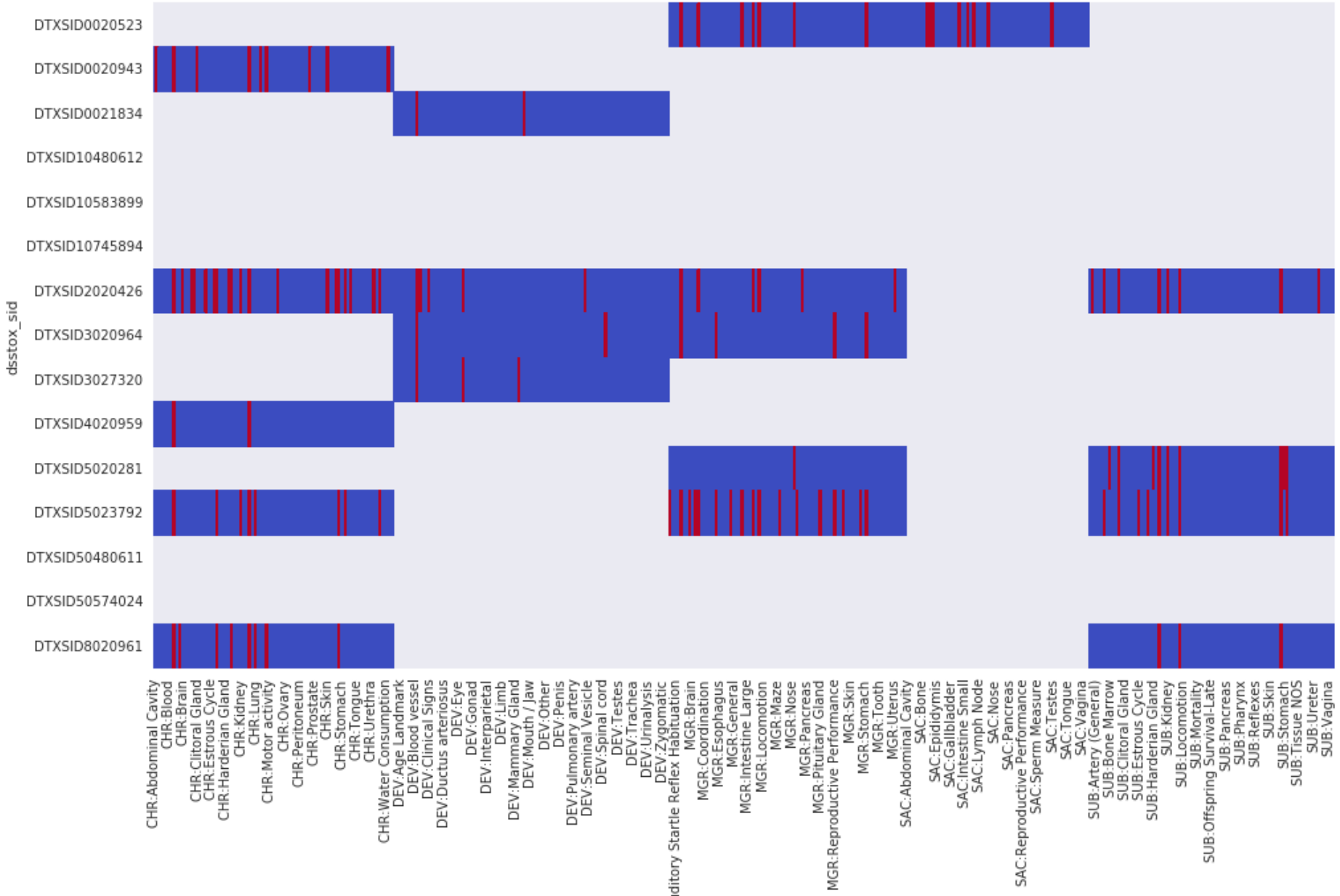
GenRA v1 in practice – step by step

Data matrix view of source analogues relative to target chemical

casrn	dsstox_sid	name	CHR:Body Weight	CHR:Bone	CHR:Bone Marrow	CHR:Clinical Chemistry	CHR:Clinical Signs	CHR:Kidney	CHR:Liver	CHR:Locomotion	CHR:Lung
99-30-9	DTXSID2020426	Dicloran	71.000 mg/kg/day	no_effect	no_effect	11.300 mg/kg/day	75.000 mg/kg/day	71.000 mg/kg/day	11.300 mg/kg/day	no_effect	no_effect
100-01-6	DTXSID8020961	4-Nitroaniline	9.000 mg/kg/day	no_effect	3.000 mg/kg/day	no_effect	no_effect	no_effect	1.500 mg/kg/day	no_effect	9.000 mg
99-55-8	DTXSID4020959	2-Methyl-5-nitroaniline	2.500 mg/kg/day	no_effect	no_effect	no_effect	no_effect	no_effect	345.000 mg/kg/day	no_effect	no_effect
98-95-3	DTXSID3020964	Nitrobenzene									
99-59-2	DTXSID0020943	2-Methoxy-5-nitroaniline	200.000 mg/kg/day	no_effect	no_effect	no_effect	no_effect	no_effect	600.000 mg/kg/day	no_effect	no_effect
100-00-5	DTXSID5020281	1-Chloro-4-nitrobenzene									
100-02-7	DTXSID0021834	4-Nitrophenol									
99-99-0	DTXSID5023792	4-Nitrotoluene	60.000 mg/kg/day	no_effect	no_effect	no_effect	no_effect	55.000 mg/kg/day	110.000 mg/kg/day	no_effect	155.000 m
824-78-2	DTXSID3027320	Sodium 4-nitrophenolate									
51-28-5	DTXSID0020523	2,4-Dinitrophenol									

Data matrix for source substances

United States Environmental Protection Agency



Look for commonality in profile across target effects

What are the most common effects across analogues

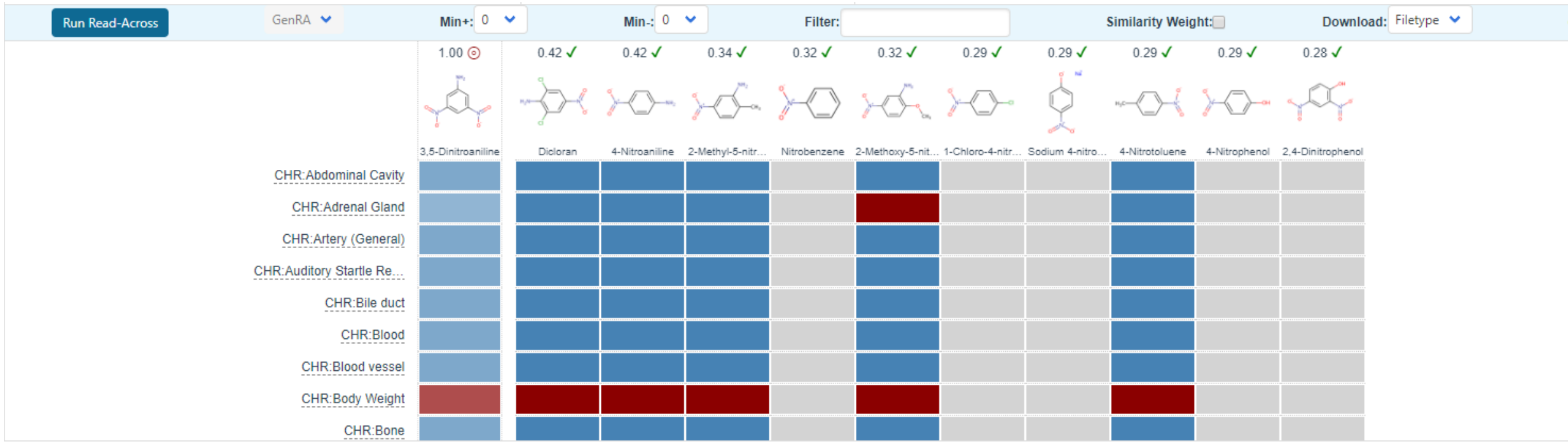
e.g. haematology, liver, kidney and spleen effects

GenRA predictions in practice: Approach

- Predictions are binary (yes/no) for toxicity effects within ToxRefDB v1 studies.
- Predictions are summarised on a study level basis where red = "positive", blue = "negative".

GenRA v1 in practice – step by step

Updated Data matrix view with GenRA predictions for target chemical



GenRA predictions in practice: Approach

a_p	a_s	a_t	auc	dsstox_sid	fp	k0	n_neg	n_pos	out	p_val	pred	s0	t0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	MGR:Offs	0.6	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	MGR:Mort	0.66	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	MGR:Sple	0.7	Neg	0.05	0
0	0.296		0	DTXSID0044151	chm_mrgn	10	3	1	SUB:Urina	0.7	Neg	0.05	0
0	0.318		0	DTXSID0044151	chm_mrgn	10	3	1	DEV:Bone	0.7	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	DEV:Mort	0.71	Neg	0.05	0
0	0.205		0	DTXSID0044151	chm_mrgn	10	3	1	SUB:Bone	0.71	Neg	0.05	0
0	0.296		0	DTXSID0044151	chm_mrgn	10	3	1	SUB:Adre	0.71	Neg	0.05	0
0	0.318		0	DTXSID0044151	chm_mrgn	10	3	1	MGR:Vagi	0.71	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	DEV:Mate	0.73	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_mrgn	10	3	1	MGR:Brair	0.73	Neg	0.05	0
0	0.22		0	DTXSID0044151	chm_r								
0	0.318		0	DTXSID0044151	chm_r								
1	0.707		0	DTXSID0044151	chm_r								
0	0.22		0	DTXSID0044151	chm_r								
0	0.236		0	DTXSID0044151	chm_r								
0	0.22		0	DTXSID0044151	chm_r								
0	0.205		0	DTXSID0044151	chm_r								
0	0.318		0	DTXSID0044151	chm_r								
0	0.22		0	DTXSID0044151	chm_r								
0	0.236		0	DTXSID0044151	chm_r								
0	0.176		0	DTXSID0044151	chm_r								
0	0.22		0	DTXSID0044151	chm_r								
0	0.22		0	DTXSID0044151	chm_r								
1	0.538		0	DTXSID0044151	chm_r								
0	0.318		0	DTXSID0044151	chm_mrgn	10	3	1	DEV:Clinic	0.78	Neg	0.05	0
1	0.538		0	DTXSID0044151	chm_mrgn	10	2	2	MGR:Kidn	0.78	Pos	0.05	0.234869
0	0.176		0	DTXSID0044151	chm_mrgn	10	4	1	CHR:Adre	0.78	Neg	0.05	0
0	0.176		0	DTXSID0044151	chm_mrgn	10	4	1	CHR:Clito	0.78	Neg	0.05	0
0	0.176		0	DTXSID0044151	chm_mrgn	10	4	1	CHR:Zymb	0.79	Neg	0.05	0
0	0.205		0	DTXSID0044151	chm_mrgn	10	3	1	SUB:Food	0.79	Neg	0.05	0
0	0.236		0	DTXSID0044151	chm_mrgn	10	4	1	CHR:Clinic	0.79	Neg	0.05	0

Considerations

Rank predictions based on the p-val and AUC values (where we have more confidence in the predictions)

Rank based on the target organ effects observed for the source analogues

GenRA – Ongoing research

- Summarising and aggregating the toxicity effect predictions to guide end users – what are the effects to be concerned about and which effect predictions are we most confident about
- Consideration of other information to define and refine the analogue selection – e.g. physicochemical similarity, metabolic similarity, reactivity similarity…
  - EPA New Chemical Categories
  - Quantifying the impact of physicochemical similarity on read-across performance (paper published)
  - Quantifying the impact of reactivity similarity on read-across performance (manuscript in late stages of development)

GenRA – Ongoing research

- Dose response information to refine scope of prediction beyond binary outcomes
  - Transitioning from qualitative to quantitative predictions – how to apply and interpret GenRA in screening level hazard assessment
    - Starting with quantitative data – e.g. acute rat oral toxicity, ToxRefDB v2 (2 manuscripts in review)

GenRA & Physchem Similarity Context

- Important context of similarity in read-across
- Models "bioavailability"
- Properties selected: Lipinski Rule of 5 (LogP, MW, # HB donors/acceptors)
- Two approaches investigated as a means to identify source analogs and evaluate their predictive performance relative to GenRA:

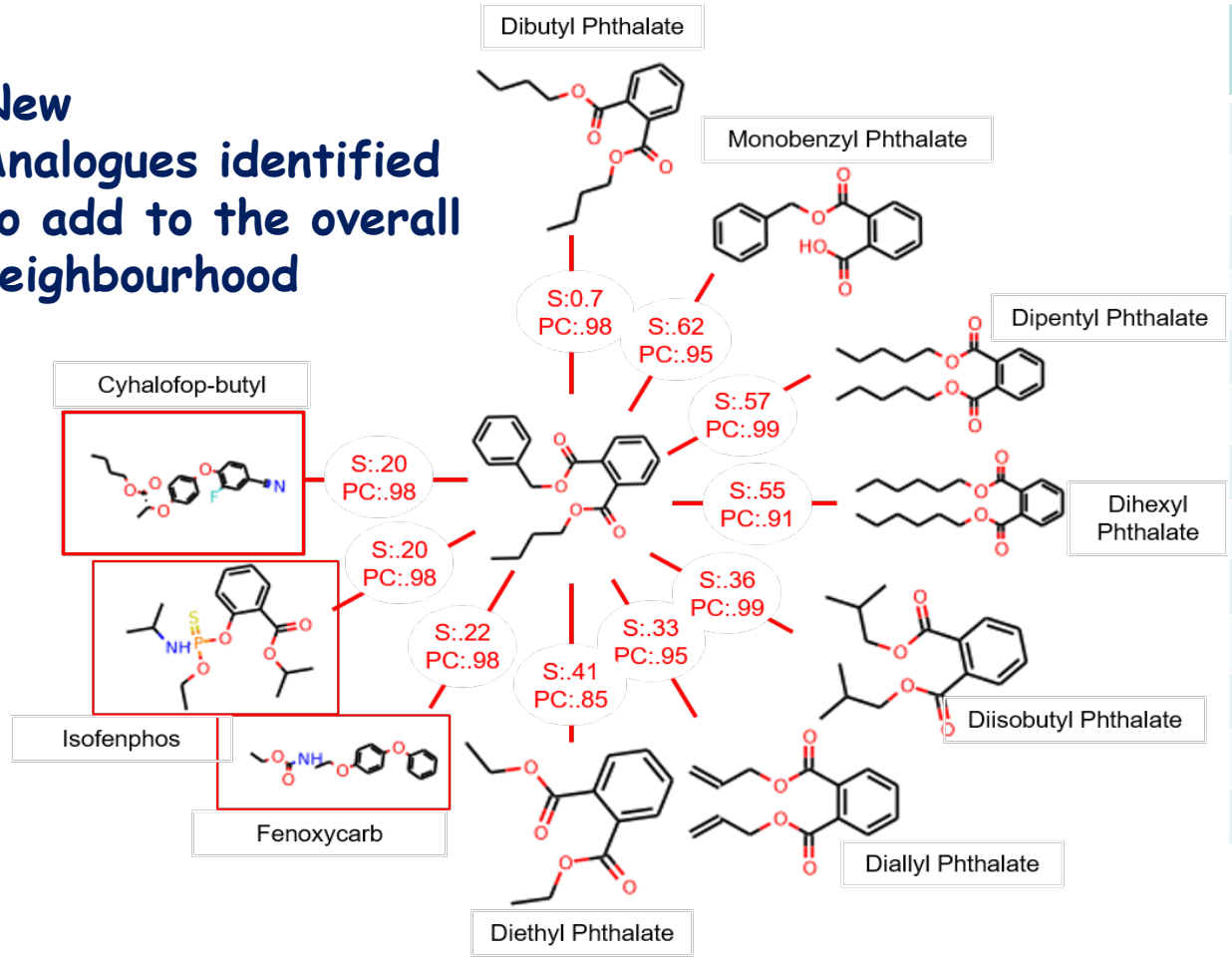
Approach 1: "Filter"	Approach 2: "Search Expansion"
Subcategorise from a set of analogues identified based on structural similarity	"Frontload" both structure and physchem into analogue identification
'Common' approach	'Novel' approach

Helman et al., 2018

Case Study: Butyl Benzyl Phthalate

Approach 2: Search Expansion

New Analogues identified to add to the overall neighbourhood



Endpoint	Baseline Prediction	Structure + Pchem Prediction
Body Weight	.78	.79
Clinical Chemistry	.27	.60
Food Consumption		
Hematology		
Kidney		
Liver		
Mortality		
Pancreas		
Prostate		
Skin	.27	.21
Spleen	0	.20
Tissue NOS	0	0
Urinary Bladder	0	0

- Adding phys-chem to similarity search overturns incorrect predictions for 2 endpoints
- Improves many others

ELSEVIER

Computational Toxicology

Available online 23 July 2018

In Press, Corrected Proof

Extending the Generalised Read-Across approach (GenRA): A systematic analysis of the impact of physicochemical property information on read-across performance

George Helman [a, b], Imran Shah [b], Grace Patlewicz [b]

Show more

https://doi.org/10.1016/j.comtox.2018.07.001

Highlights

- GenRA approach is summarised in the context of the category workflow.
- The impact of physicochemical information on read-across performance was assessed in 2 ways: filtering and search expansion.
- Search expansion resulted in an up to 9% improvement in read-across performance for 10 of the 50 data rich target organs.
- Results are summarised on a neighbourhood (chemical category) basis.
- A case study substance is used to compare and contrast the read-across performance using the 2 approaches.

Refinements to the GenRA approach

- Transitioning GenRA from binary predictions to quantitative predictions
- Investigated extending GenRA using the acute oral rat systemic toxicity data collected as part of the ICCVAM Acute toxicity workgroup
- NICEATM-NCCT effort to collate a large dataset of acute oral toxicity to evaluate the performance of existing predictive models and investigate the feasibility of developing new models

Refinements to the GenRA approach: Acute toxicity

Database Resource	Rows of Data (number of LD50 values)	Unique CAS
ECHA (ChemProp)	5533	2136
JRC AcutoxBase	637	138
NLM HSDB	4082	2238
OECD (eChemPortal)	10206	2314
PAI (NICEATM)	364	293
TEST (NLM ChemIDplus)	13689	13545

Rat oral LD50s:
16,297 chemicals total
34,508 LD50 values

Require unique LD50 values with mg/kg units

15,688 chemicals total
21,200 LD50 values

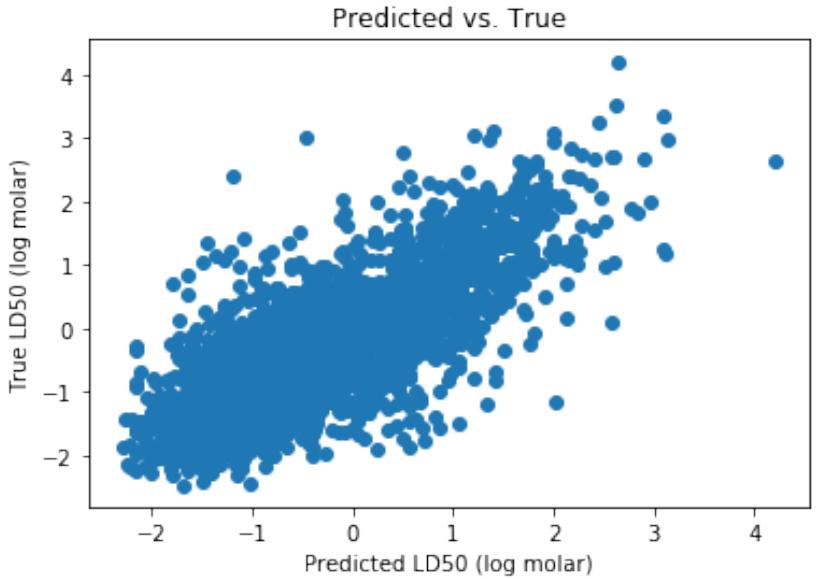
Preprocessing for modelling

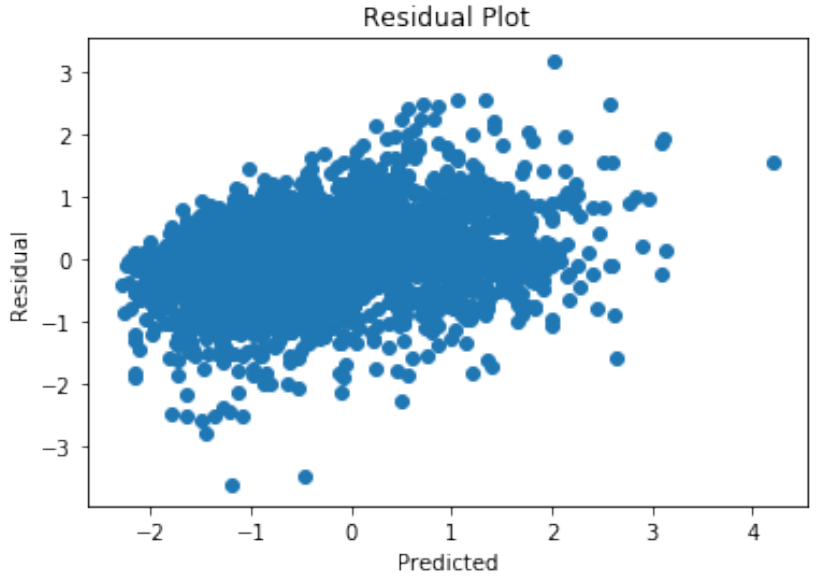
11,992 chemicals
16,209 LD50 values

Karmaus et al, 2018; Kleinstreuer et al., 2018

Refinements to the GenRA approach: Acute toxicity

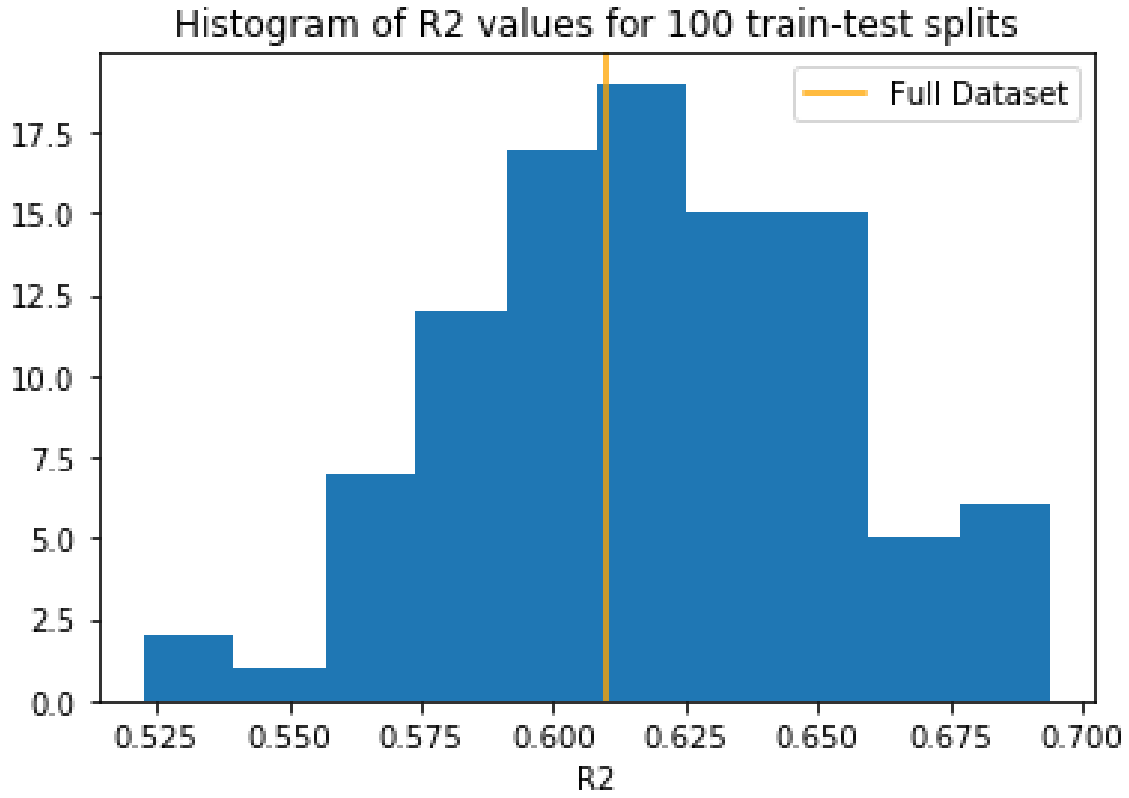
- Search for a maximum of 10 nearest neighbours on entire dataset
- Use a min similarity threshold of 0.5





- R^2 = 0.61
- RMSE = 0.58
- A few outliers, but not too extreme
- Residuals clustered around zero with no obvious patterns

Refinements to the GenRA approach: Acute toxicity



- 75-25 train-test splits
- R^2 values range from 0.52 to 0.69
- GenRA performs strongly and robustly on this acute tox data set.

Helman et al., in review

Take home messages

- Harmonised framework for read-across provides opportunities for NAM data
- GenRA developed is aligned with this framework
- Illustrated how GenRA baseline can been applied in practice
- Highlight ongoing research in extending the approach
  - quantitative impact of physicochemical similarity (as it relates to bioavailability)
  - transitioning to quantitative predictions of PODs

Acknowledgements

- Many but in particular..
- NCCT
- Imran Shah
- George Helman
- Tony Williams
- Jeff Edwards
- Jason Lambert

- NCEA
- Lucy Lizarraga