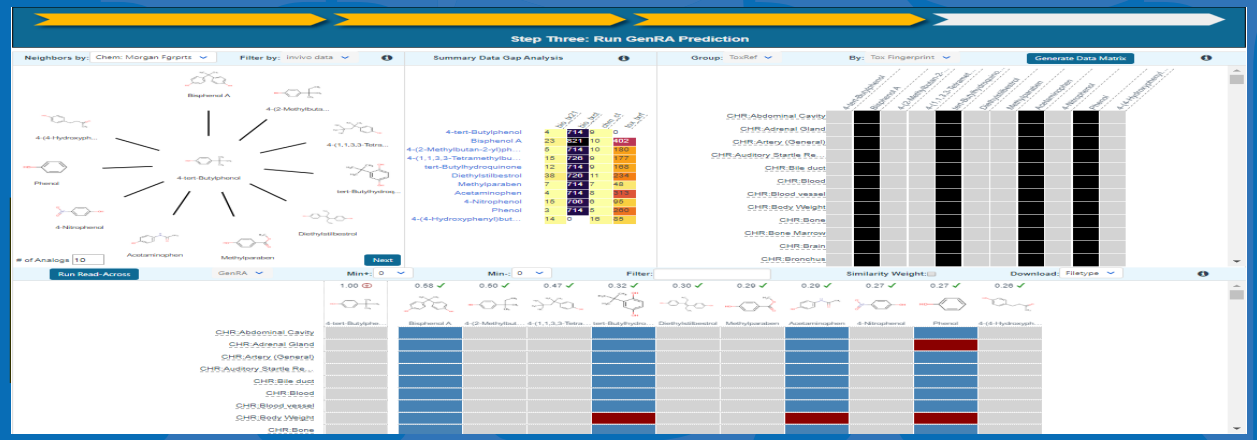
United States Environmental Protection Agency

Building Scientific Confidence in the Development and Application of Objective Read-across Approaches



Grace Patlewicz
Center for Computational Toxicology & Exposure (CCTE), US EPA

Acknowledgements

- Imran Shah – co-lead on Generalised Read-across (GenRA)
- George Helman (former student)
- Tia Tate
- Willysha Jenkins

EPA United States Environmental Protection Agency

Outline

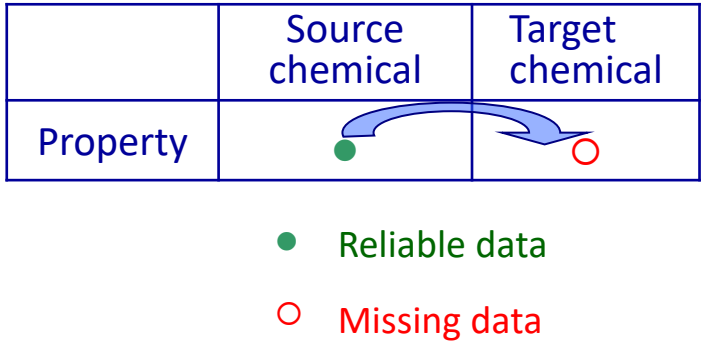
- Read-across - definition
- Background context – tools, frameworks

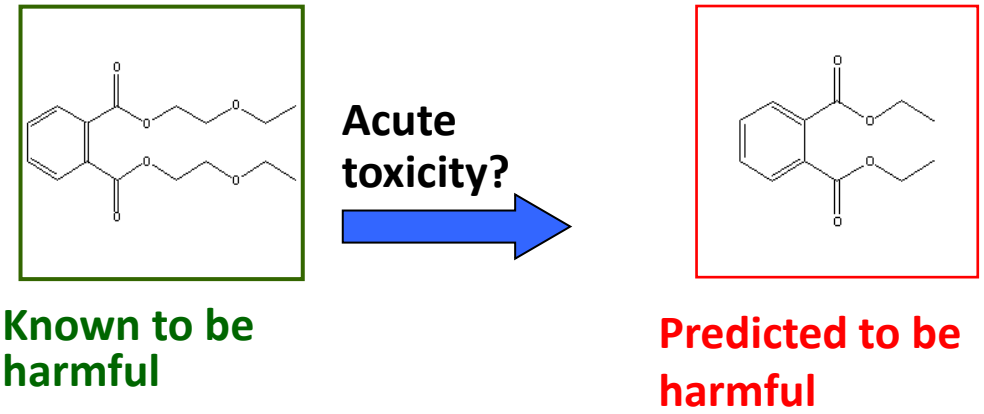
- Generalised Read-across (GenRA): A data driven approach to read-across
  - Implementations of GenRA
  - Recent work applying GenRA

- ICCVAM Read-Across Workgroup activities

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.





Ongoing issues with read-across

- Although there is much guidance for developing read-across assessment, acceptance remains an issue, not helped since read-across still remains a subjective, expert driven assessment.
- One issue thwarting acceptance relates to the "uncertainty of the read-across prediction".
- As such there have been many efforts to identify the sources of uncertainty in read-across, characterise them in a consistent manner and identify practical strategies to address and reduce those uncertainties.
- Notable in these efforts have been the development of frameworks for the assessment of read-across, evaluating the utility of New Approach Methods (NAMs).
- Quantifying uncertainty and performance of read-across is still a need as are ways to better characterise different similarity contexts (metabolism, reactivity etc.)

5

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 neighbours

ABSTRACT

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

Published by

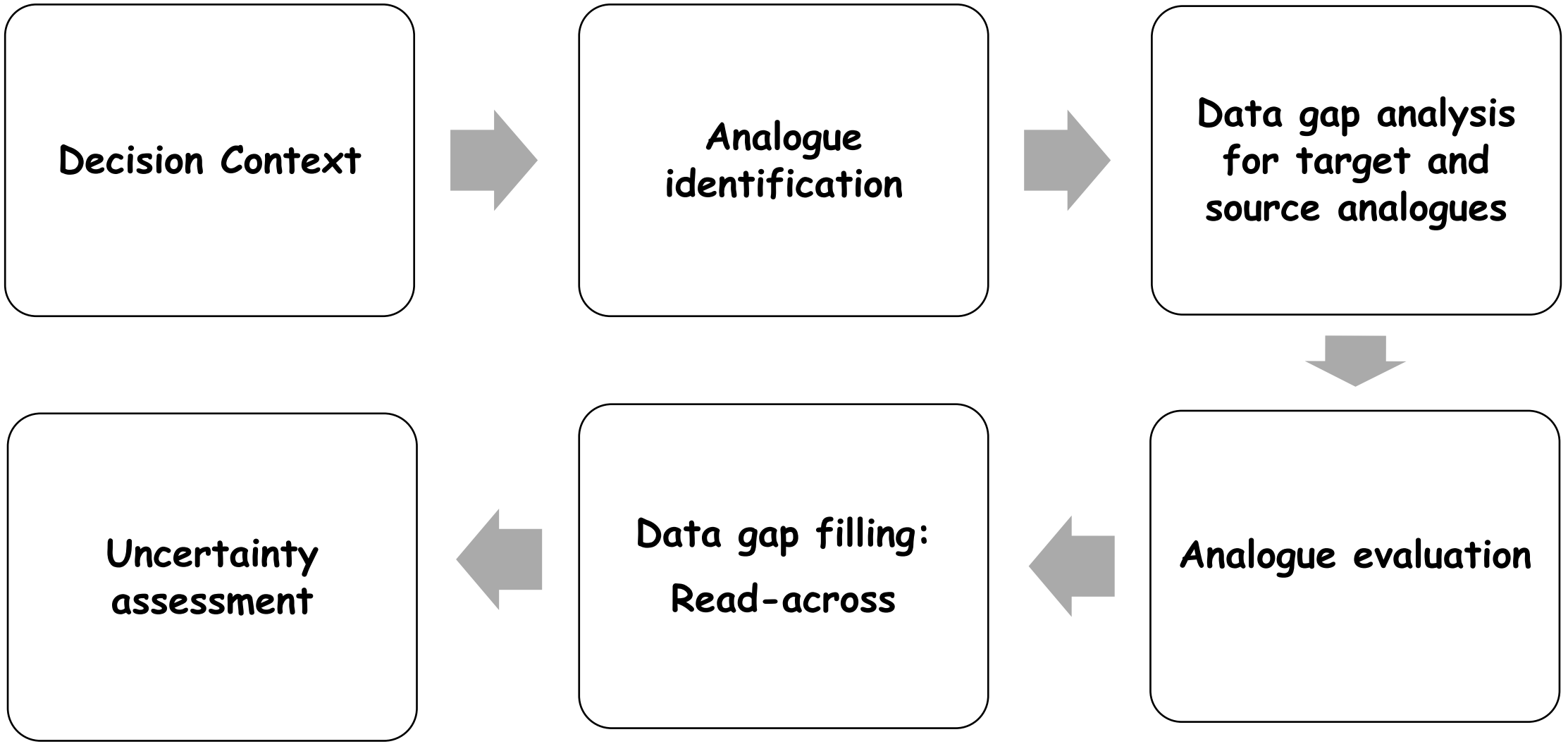
(Patlewicz et al., 2017)

Summary of key features of selected publicly available read-across tools.

	AIM	ToxMatch	Ambit	OECD Toolbox	CBRA	ToxRead	CIIPro
Development timeline	Java based version is dated 2012. Initial development of web version was 2005.	First public version released in Dec 2006	Original AMBIT tool was developed in 2004–2005	Proof of concept released in 2008	Implementation of the Low et al. [27]	Implementation of Gini et al. [22]	Implementation described in Russo et al. [45]
Type of Tool	Standalone	Standalone	Web-based and standalone	Standalone or Client/Server	Standalone	Standalone	Web-based
Latest Version	1.01 (Nov 2013) Static	1.07 (Jan 2009) Static	3.0.3 Ongoing Enhanced in 2013–2015	3.4 (July 2016) Version 4 released April 2017 Ongoing	0.75 First release	0.11 BETA Ongoing	First release
Developed by	SRC Inc	Ideaconsult Ltd	Ideaconsult Ltd	LMC, Bourgas	Fourches Lab at North Carolina State University	Istituto di Ricerche Farmacologiche Mario Negri	Zhu Research Group at Rutgers University
Available from	https://www.epa.gov/tsca-screening-tools/analog-identification-methodology-aim-tool	https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/toxmatch	http://cefic-lri.org/lri_toolbox/ambit/	www.qsartoolbox.org	http://www.fourches-laboratory.com/software	http://www.toxread.eu/	http://ciipro.rutgers.edu/
Accepted Chemical Input	CAS, Name, SMILES, structure drawing/import	CAS, Name, SMILES, InChI	Name, identifiers, SMILES, InChI	CAS, Name, SMILES, structure drawing, MOL, sdf	Mol file, descriptors as txt	SMILES	PubChem CID, CAS, IUPAC, SMILES, InChI
Endpoint Coverage	N/A	Any based on user input	IUCLID[a] 5-supported endpoints (43 total)	Any as per the regulatory endpoints	Any based on user input	Mutagenicity and Bioconcentration Factor (BCF)	Any based on user input
Analogue Identification Approach	Fragment matching	Distance and correlation based similarity indices based on descriptors or fingerprints	Substructure or similarity searching using structure, name, SMILES, InChI	Category definition followed by subcategorisations	Tanimoto distance using chemical and biological descriptors	VEGA similarity algorithm	Weighted Estimated Biological Similarity
Neighbour Selection	Automatic	Automatic	Manual	Automatic + Manual Filter	Automatic	Automatic	Automatic + Manual Filter
Data Source	Tool provides inventory index	User provided or tool provided	User and tool provided	User provided or tool provided	User provided	Tool provided as a result of the EU ANTARES project	User provided but tool provides PubChem in vitro data
Quantitative vs Qualitative	N/A	Both	User determined - Qualitative	Both	Qualitative	Qualitative for mutagenicity, quantitative for BCF	Qualitative
Visualisation	None	Standard 2D plots, histograms and similarity matrix	None	Standard 2D Plots	Radial plot of neighbours	Interactive Neighbour plot	Activity Plot
Output/Export	Output reports in the form of HTML, pdf or Excel	sdf or txt files of data, image files of plots	Assessment report as docx or xlsx, data matrix as xlsx	IUCLID format, pdf and rtf files of prediction report, text files of data, image files of plots etc	NA	Image file of plot	Tabulation of predictions and image of similarity plot

[a] IUCLID stands for International Uniform Chemical Information Database. IUCLID is a software program for the administration of data on chemical substances first developed to fulfill EU information requirements under REACH.

Read-across workflow



A harmonised hybrid read-across workflow

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

- Where do NAM data fit?
- How should we transition to data-driven approaches?
- Quantifying the uncertainty in the read-across predictions made?

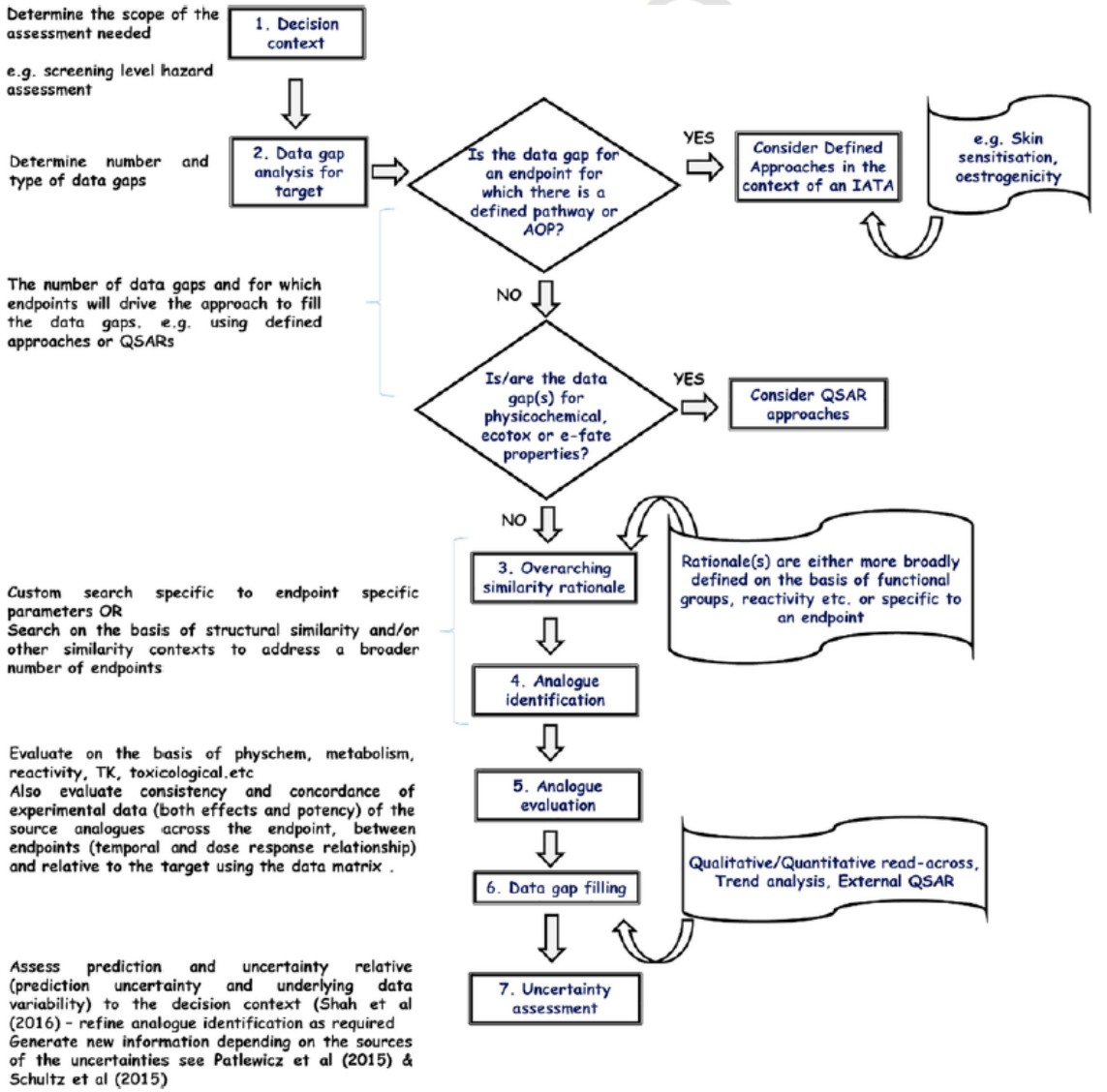


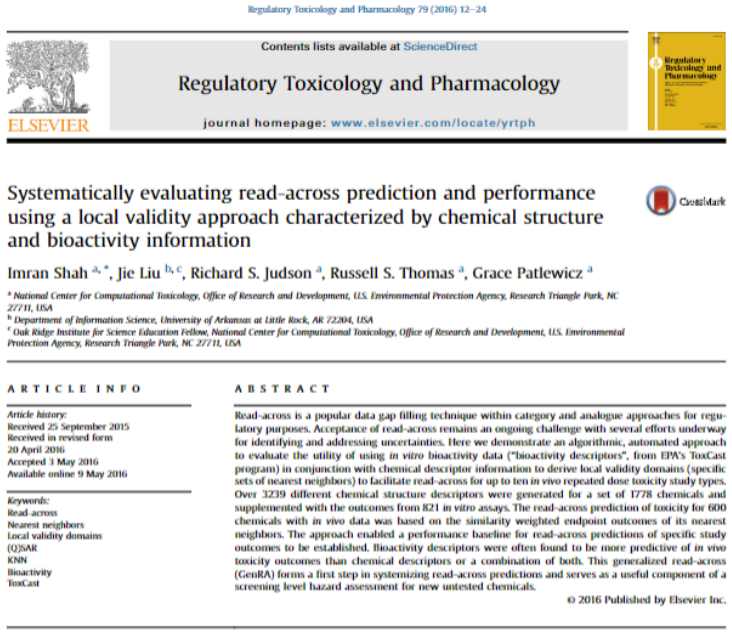
Fig. 9. A harmonised hybrid development and assessment framework.

Patlewicz et al., 2018

EPA United States Environmental Protection Agency

GenRA (Generalised Read-Across)

- Predicting toxicity as a similarity-weighted activity of nearest neighbours based on chemistry and bioactivity descriptors (Shah et al, 2016)
- Goal: To establish an objective performance baseline for read-across and quantify the uncertainty in the predictions made



Regulatory Toxicology and Pharmacology 79 (2016) 12–24

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information

Imran Shah [a, *], Jie Liu [b, c], Richard S. Judson [a], Russell S. Thomas [a], Grace Patlewicz [a]

[a] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[b] Department of Information Science, University of Arkansas at Little Rock, AR 72204, USA
[c] Oak Ridge Institute for Science Education Fellow, National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

ARTICLE INFO

Article history:
Received 25 September 2015
Received in revised form
20 April 2016
Accepted 3 May 2016
Available online 9 May 2016

Keywords:
Read-across
Nearest neighbors
Local validity domains
(Q)SAR
KNN
Bioactivity
ToxCast

ABSTRACT

Read-across is a popular data gap filling technique within category and analogue approaches for regulatory purposes. Acceptance of read-across remains an ongoing challenge with several efforts underway for identifying and addressing uncertainties. Here we demonstrate an algorithmic, automated approach to evaluate the utility of using *in vitro* bioactivity data ("bioactivity descriptors", from EPA's ToxCast program) in conjunction with chemical descriptor information to derive local validity domains (specific sets of nearest neighbors) to facilitate read-across for up to ten *in vivo* repeated dose toxicity study types. Over 3239 different chemical structure descriptors were generated for a set of 1778 chemicals and supplemented with the outcomes from 821 *in vitro* assays. The read-across prediction of toxicity for 600 chemicals with *in vivo* data was based on the similarity weighted endpoint outcomes of its nearest neighbors. The approach enabled a performance baseline for read-across predictions of specific study outcomes to be established. Bioactivity descriptors were often found to be more predictive of *in vivo* toxicity outcomes than chemical descriptors or a combination of both. This generalized read-across (GenRA) forms a first step in systemizing read-across predictions and serves as a useful component of a screening level hazard assessment for new untested chemicals.

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

Jaccard similarity:

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

$\alpha \in \{chm, bio, bc\}$

$\beta \in \{bio, tox\}$

y_i = *predicted activity of chemical* (c_i)

x_j^{β} = *activity of* c_j in β

s_{ij}^{α} = *Jacccard similarity between* $\boldsymbol{x}_i^{\alpha}, \boldsymbol{x}_j^{\alpha}$

k = *up to* k *nearest neighbours*

Read-across workflow in GenRA v1.0

Decision Context

Screening level assessment of hazard based on toxicity effects from ToxRefDB v1

Analogue identification

Similarity context is based on structural characteristics

Data gap analysis for target and source analogues

Analogue evaluation

Evaluate consistency and concordance of experimental data of source analogues across and between endpoints

Read-across

Similarity weighted average – many to one read-across

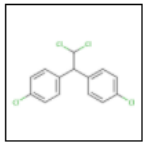
Uncertainty assessment

Assess prediction and uncertainty using AUC and p value metrics

GenRA tool in reality

- GenRA v1.0 Integrated into the EPA CompTox Chemicals Dashboar

GenRA tool in practice



p,p'-DDD
72-54-8 | DTXSID4020373
Searched by DSSTox Substance Id.

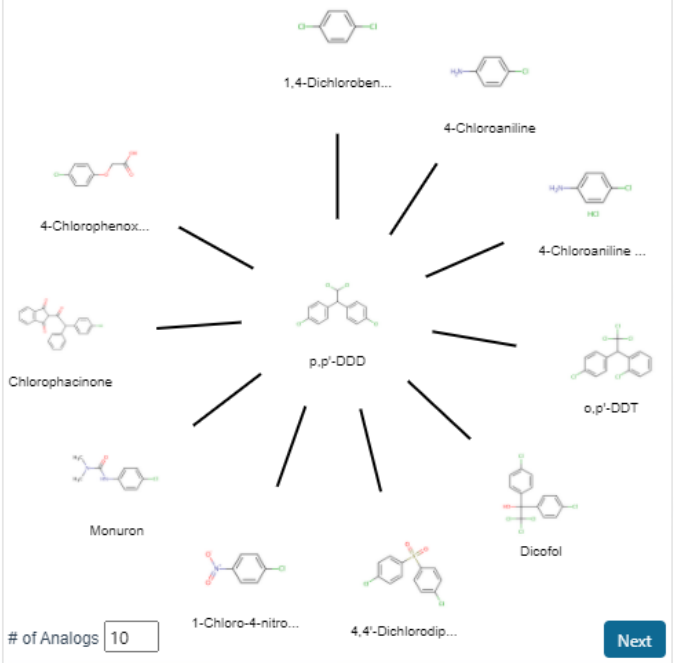
Search for a chemical and click on the GENRA link on the lefthand panel

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

Generalized Read-Across (GenRA)

Step One: Analog Identification and Evaluation

Neighbors by: Chem: Morgan Fgrprts Filter by: invivo data



# of Analogs 10 Next

Radial plot with Target chemical of interest in the centre and source analogues (similar) ordered clockwise by decreasing similarity (Jaccard)

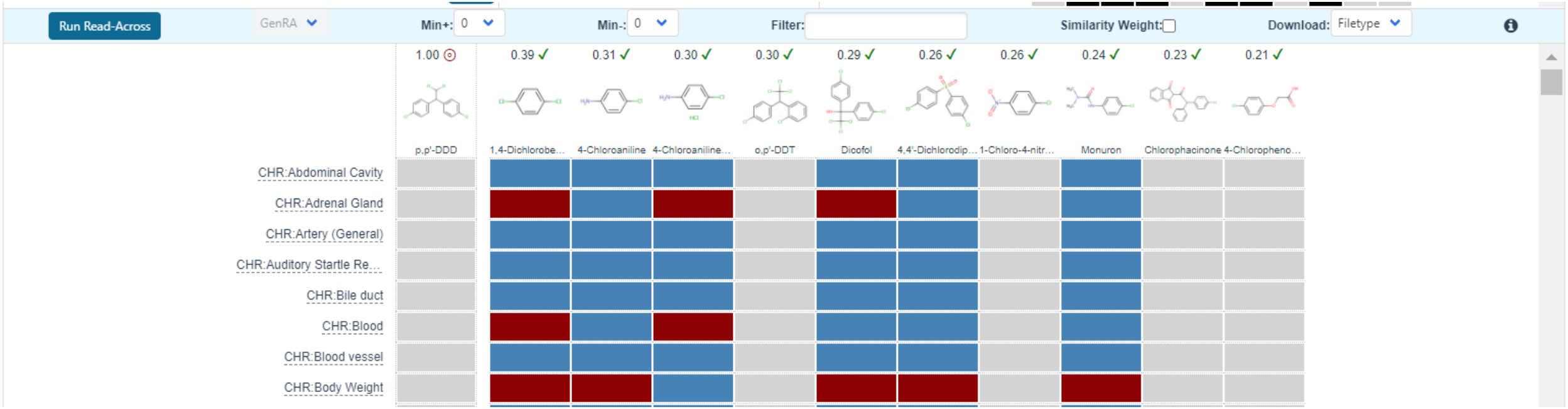
GenRA tool in practice

- How data poor is my target and what data exists for the source analogues identified
- Do they address the data gaps of interest for the target chemical?

GenRA tool in practice

What is the consistency and concordance across my source analogues?
Should I deselect analogues from consideration from the entire set of predictions?
Should I consider subcategorising the analogues selected?

Toxicity data represented as binary outcomes – red (positive), blue (negative), grey (no data)



United States Environmental Protection Agency

GenRA tool in practice

First column is updated with predictions

GenRA Tool in practice

- Database underpinning GenRA v1.0: ToxRefDB v1
  - Different study types and effects within them are predicted e.g. chronic_liver is annotated as CHR_liver
  - Negative results – assume that if a particular guideline study was conducted but the effects were not reported than a chemical would be negative for that particular effect for that type of guideline study
  - Positive results – min dose at which toxicity effects are observed in a study
- Prediction: Similarity weighted activity
- Performance is categorised by the AUC of the ROC
  - The significance was empirically estimated by constructing a null distribution by permuting the toxicity values 100 times and calculating the fraction of times the AUC was more extreme than what would be observed by chance (this is reported as the p-value).

GenRA Tool in practice

- Ability to export the predictions as an excel file
- Output can be analysed in different ways

role	target		analog
preferred name	p,p'-DDD		1,4-Dichlorobenzene
dsstox_sid	DTXSID4020373		DTXSID1020431
molecular weight	320.03		147
similarity	1		0.391304348
CHR:Abdominal Cavity	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Adrenal Gland	GenRA Pos Act=1 (0.546) AUC=0 p=0.975		600mg/kg/day
CHR:Artery (General)	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Auditory Startle Reflex Habituation	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Bile duct	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Blood	GenRA Neg Act=0 (0.386) AUC=0 p=0.95		150mg/kg/day
CHR:Blood vessel	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Body Weight	GenRA Pos Act=1 (0.832) AUC=0 p=0.8		300mg/kg/day
CHR:Bone	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Bone Marrow	GenRA Neg Act=0 (0.168) AUC=0 p=0.85		no_effect
CHR:Brain	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect
CHR:Bronchus	GenRA Neg Act=0 (0) AUC=0 p=1		no_effect

alog		analog		analog		analog
l-Dichlo		4-Chloroa		4-Chloroa		o,p'-DDT
XSID102		DTXSID902		DTXSID402		DTXSID602
147		127.57		164.03		354.48
391304		0.310345		0.3		0.295455
_effect		no_effect		no_effect		no_data
CHR:Adrer	GenRA Pos	600mg/kg/		no_effect		18mg/kg/c
CHR:Arter	GenRA Ne	no_effect		no_effect		no_data
CHR:Audit	GenRA Ne	no_effect		no_effect		no_data
CHR:Bile d	GenRA Ne	no_effect		no_effect		no_data
CHR:Blooc	GenRA Ne	150mg/kg/		no_effect		3mg/kg/da
CHR:Blooc	GenRA Ne	no_effect		no_effect		no_data
CHR:Body	GenRA Pos	300mg/kg		375mg/kg		no_effect

EPA United States Environmental Protection Agency

GenRA Tool in practice

- Rank order positive results based on AUC and p values
- Look at the distribution of positive vs negatives predictions
- Explore what effects are being identified for the source analogues – consider identifying the underlying data for source analogues (elsewhere on the Dashboard) – is there a critical effect that is driving the toxicity that should be compared with the target chemical predictions?
- ……
- Depends on the decision context and the level of uncertainty that can be tolerated.

GenRA tools

- Efforts are underway to update the underlying data sources of the webapp GenRA for a summer release*

- An alternative and programmatic batch means of using GenRA is available through genra-py*, a standalone python library to enable user specific datasets to be analysed – see https://github.com/i-shah/genra-py (Shah et al, 2021)

Bioinformatics, 2021, 1–2
doi: 10.1093/bioinformatics/btab210
Advance Access Publication Date: 27 March 2021
Applications Note

Data and text mining

Generalized Read-Across prediction using genra-py

Imran Shah *, Tia Tate and Grace Patlewicz

Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, NC 27709, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 14, 2020; revised on March 15, 2021; editorial decision on March 24, 2021; accepted on March 25, 2021

Abstract

Motivation: Generalized Read-Across (GenRA) is a data-driven approach to estimate physico-chemical, biological or eco-toxicological properties of chemicals by inference from analogues. GenRA attempts to mimic a human expert's manual read-across reasoning for filling data gaps about new chemicals from known chemicals with an interpretable and automated approach based on nearest-neighbors. A key objective of GenRA is to systematically explore different choices of input data selection and neighborhood definition to objectively evaluate predictive performance of automated read-across estimates of chemical properties.

Results: We have implemented genra-py as a python package that can be freely used for chemical safety analysis and risk assessment applications. Automated read-across prediction in genra-py conforms to the scikit-learn machine learning library's estimator design pattern, making it easy to use and integrate in computational pipelines. We demonstrate the data-driven application of genra-py to address two key human health risk assessment problems namely: hazard identification and point of departure estimation.

Availability and implementation: The package is available from github.com/i-shah/genra-py.

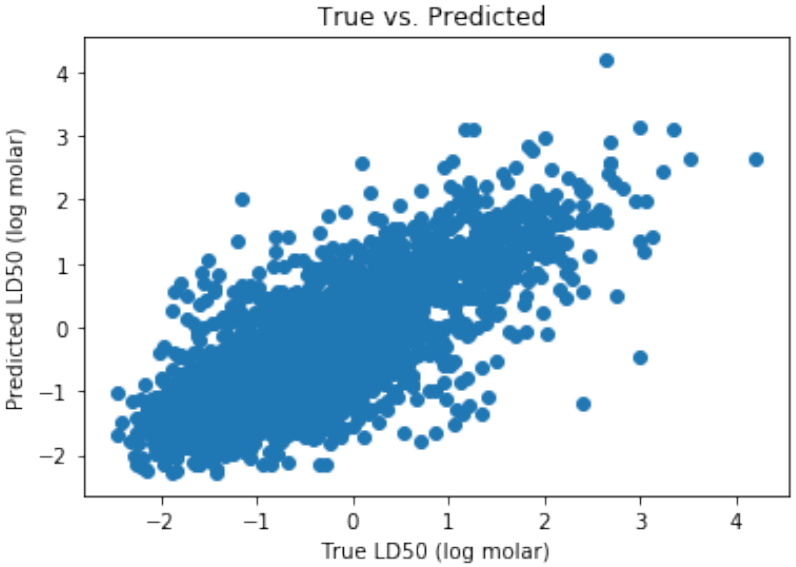
Contact: shah.imran@epa.gov

GenRA – Current research

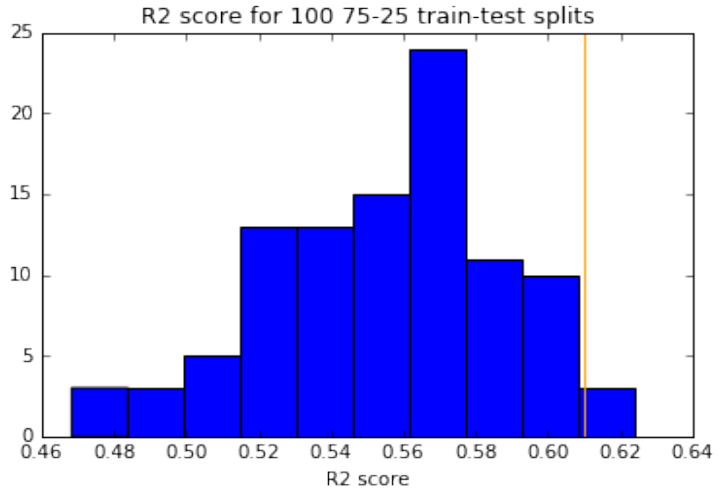
- Consideration of other information to define and refine the analogue selection & evaluation
  - physicochemical similarity (Helman et al 2018)
  - metabolic similarity (Patlewicz *in prep*),
  - reactivity similarity (Nelms et al 2018)
  - transcriptomics similarity (Tate et al, *under review*)*
- Transitioning to quantitative predictions of toxicity
  - Using GenRA to predict LOAEL, acute oral LD50 (Helman et al 2019a,b)
- Developing a compendium of expert driven read-across examples to investigate how data driven read-across with NAM data can mirror expert assessments (Jenkins et al *in prep*)*

Acute oral toxicity : 'Global' performance

- Search for a maximum of 10 nearest neighbours on entire dataset on the basis of Morgan chemical fingerprints
- Use a min similarity threshold of 0.5



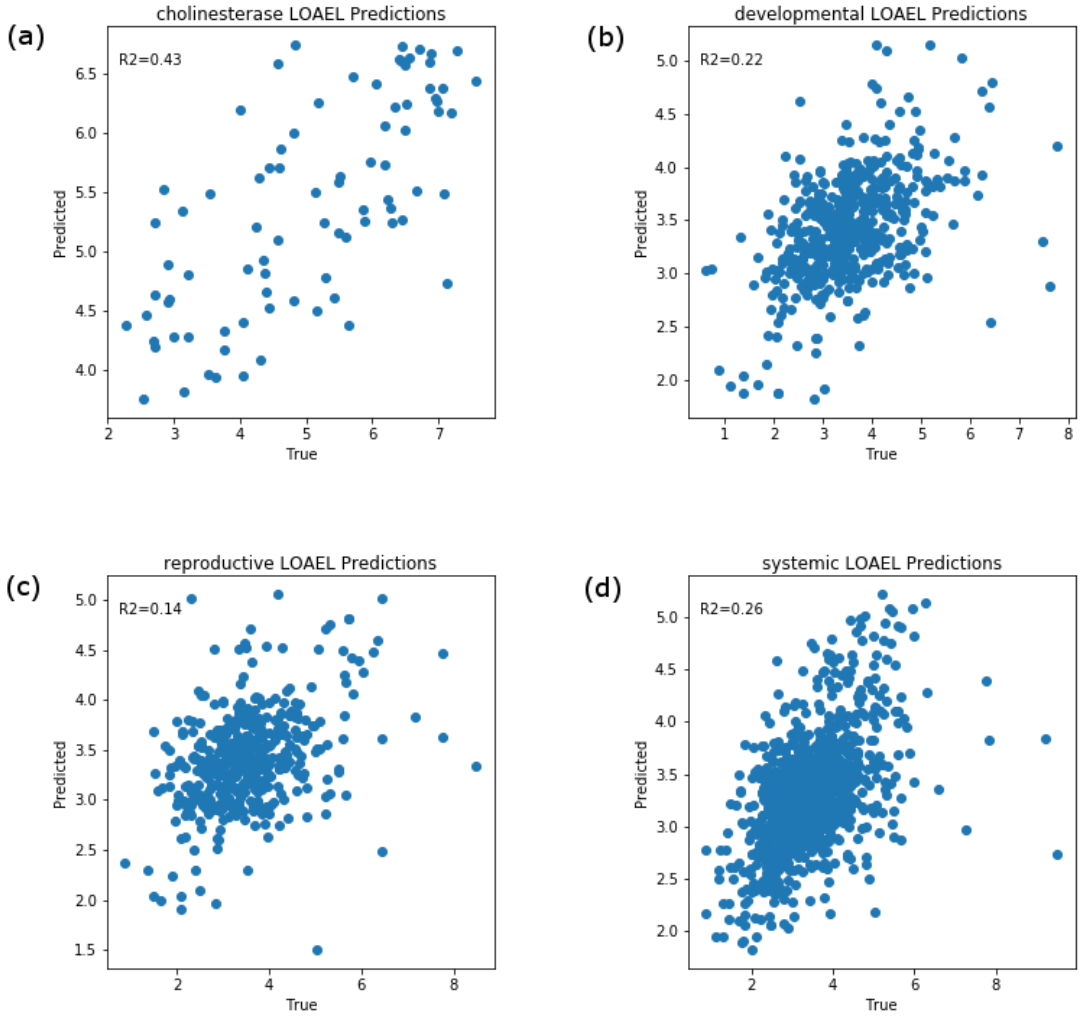
- Linear regression used to fit predicted and observed LD50 values
- R^2 = 0.61
- RMSE = 0.58



- Monte Carlo CV
- Estimate confidence in R2
- 75-25 train-test splits
- R^2 values range from 0.46 to 0.62

Helman et al., 2019a

LOAEL prediction : 'Global' performance



GenRA Predictions using Morgan fingerprints with k=10 and s=0.05 (mean aggregated LOAELs)
Linear regression used to fit predicted and observed LOAEL values

Endpoint Category	R2
Cholinesterase	0.43
Developmental	0.22
Reproductive	0.14
Systemic	0.26

Helman et al., 2019b

Characterising metabolic similarity

	Parent_DTXSID	Frag	Parent_smiles	Metabolite_smiles
0	DTXSID20375106	[#6](=[#8])(-[#8])-[#6]>>[#6]	O=C(O)C(F)(F)OC(F)(F)C(F)(F)OC(F)(F)C(=O)O	O=C(O)C(F)(F)OC(F)(F)C(F)(F)OC(F)F
1	DTXSID7027831	[#6]-[#7]>>[#7]	CN(CCO)S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(...	O=S(=O)(NCCO)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(...
2	DTXSID7027831	[#6]>>[#8]=[#6]	CN(CCO)S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(...	CN(CC(=O)O)S(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(...
3	DTXSID7027831	[#6](-[#6]-		F)C(F)(F)C(F)(F)C(F)(F)...
4	DTXSID8051419	[#6]-[#7+] [#8]-[#6]		C(F)(F)C(F)(F)C(F)

	metab_fp_0	metab_fp_1	metab_fp_2	metab_fp_3	metab_fp_4	metab_fp_5	meta
DTXSID00190950	0	0	0	0	0	0	0
DTXSID00192353	0	0	0	0	0	0	0
DTXSID00194615	0	0	0	0	0	1	0
DTXSID00379268	0						
DTXSID00379884	0						

	DTXSID00190950	DTXSID00192353	DTXSID00194615	DTXSID00379268	DTXSID0037
DTXSID00190950	1.0	0.0	0.0	0.0	0.0
DTXSID00192353	0.0	1.0	0.0	0.5	0.0
		)	1.0	0.0	0.0
		5	0.0	1.0	0.0
DTXSID00379884	0.0	0.0	0.0	0.0	1.0

Creating custom fingerprints to characterise metabolic transformations

GenRA – Overall goal

- Quantify the contribution that different similarity contexts play in toxicity prediction and how that differs depending on the toxicity endpoint of interest, the chemical of interest and whether it mirrors expert driven read-across

- Quantify level of confidence for prediction made

=> objective, reproducible read-across assessments

GenRA Summary

- GenRA is an attempt to move towards an objective read-across approach where uncertainties and performance can be quantified. Provides opportunities for NAM data to be incorporated.
- GenRA v1.0 establishes a baseline in performance. The approach relies on chemical descriptors to predict binary toxicity values but work continues to characterise other contexts of similarity (e.g. mechanistic, reactivity, metabolism) and quantify their contribution in predicting *in vivo* toxicity outcomes.
- GenRA v1.0 exists as an app within the Dashboard to facilitate a workflow approach to make read-across predictions. An updated version is anticipated this summer. A python package (genra-py) has been released (March 2021) to facilitate batch processing using user specific datasets.

- Items* will be presented at QSAR2021 – see qsar2021.org

ICCVAM Read-Across Workgroup

- In 2018, US Agencies established a read-across workgroup (RAWG) under ICCVAM to develop and implement a plan to build capacity in the development and application of read-across approaches and to harmonise them.
- Initially, the RAWG summarised current experiences and needs, and catalogued the different tools applied (Patlewicz et al (2019)
- More recent RAWG efforts have been focused on developing a compendium of member agency read-across case studies to inform guiding principles for different read-across decision contexts.
- Several case studies were discussed ranging from the utility of metabolic data in categories, the FDA Extended Decision Tree for TTC to the qualitative use of ToxCast data to characterise bioactivity similarity of a target and candidate analogues.
- A short manuscript is in preparation to summarise the case studies and extract any general guiding principles that will be informative as part of ongoing efforts to refine existing guidance e.g. OECD grouping guidance.