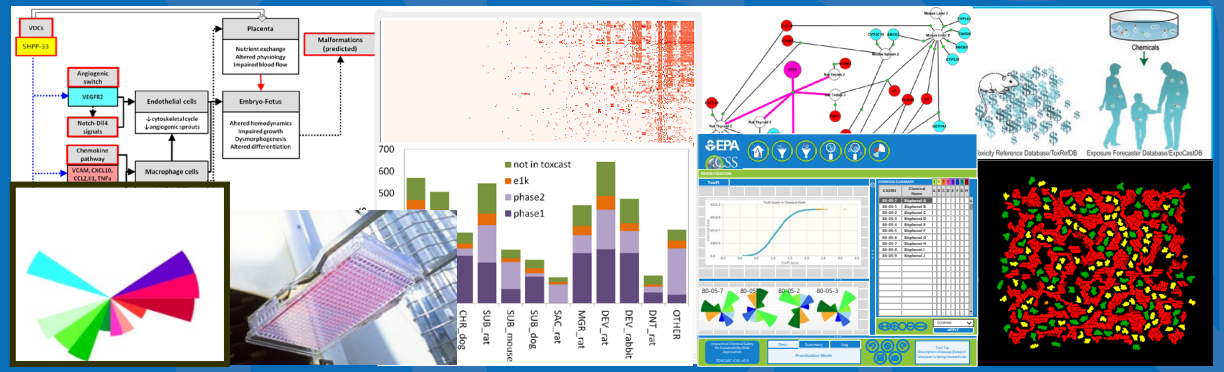
United States Environmental Protection Agency

Towards systematic read-across using Generalised Read-Across (GenRA)



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

Conflict of Interest Statement

- No conflicts of interest to declare.

Disclaimer:

- The views expressed herein are those of the presenter and do not necessarily reflect the views or policies of the U.S. EPA

Outline

- Read-Across – background, issues – quick primer
- Generalised Read-Across (GenRA)
- Summary remarks
- Acknowledgements

Acknowledgements

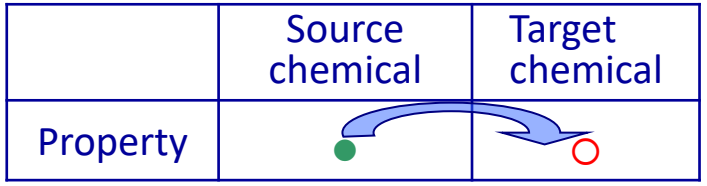
- Too many to list..
- Imran Shah (co-lead on GenRA)
- GenRA Developmental Team
- Tony Williams (especially for all the work related to the CompTox Chemicals Dashboard)
- Past* and present students
  - George Helman*
  - Mark Nelms*
  - Willysha Jenkins*
  - Tia Tate*
  - Matthew Boyce*
  - Louis Groff
  - Matthew Adams
  - Brett Hagan

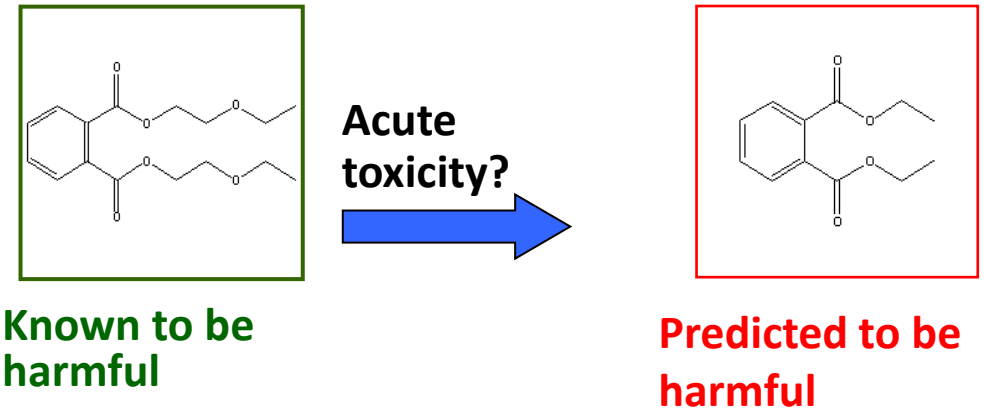
Definitions: Chemical grouping approaches

- Read-across describes one of the methods for filling data gaps in either the analogue or category approaches i.e. not to be confused with the "analogue approach"
- "Analogue approach" refers to grouping based on a very limited number of chemicals (e.g. target substance) + source substance)
- "Category approach" is used when grouping is based on a more extensive range of analogues (e.g. 3 or more members)
- A chemical category is a group of chemicals whose physico-chemical and human heath and/or environmental toxicological and/or environmental fate properties are likely to be similar or follow a regular pattern as a result of structural similarity (or other similarity characteristics).

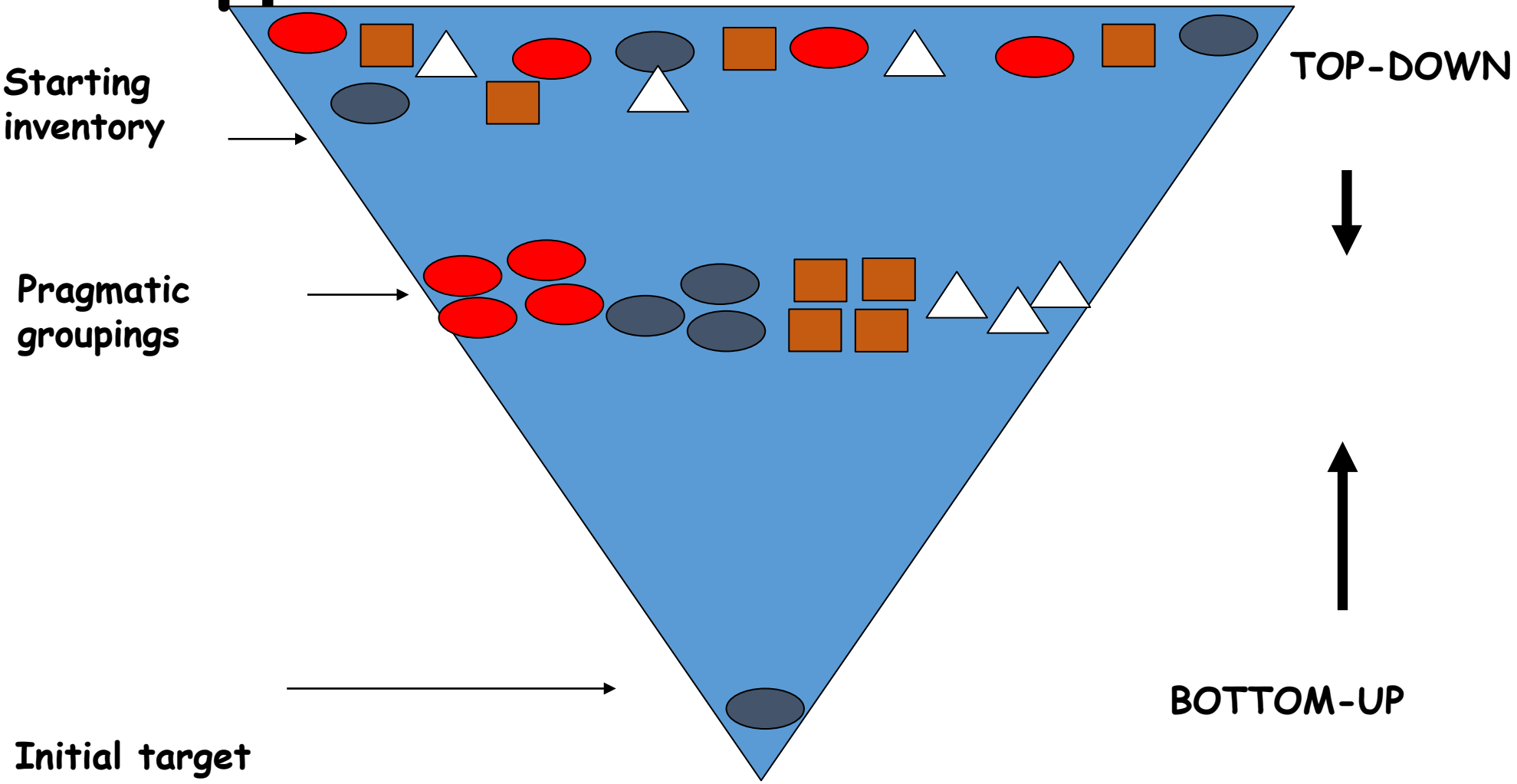
Read-across

- 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.
- Used within analogue and category approaches.
- 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.





Decision context will have a bearing on approach….



Ongoing issues with read-across

- Lots of guidance for developing read-across assessment, acceptance an issue, not helped since read-across still remains a subjective, expert driven assessment.
- One issue thwarting acceptance related to the "uncertainty of the read-across prediction".
- Many efforts undertaken 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.)

8

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.

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?

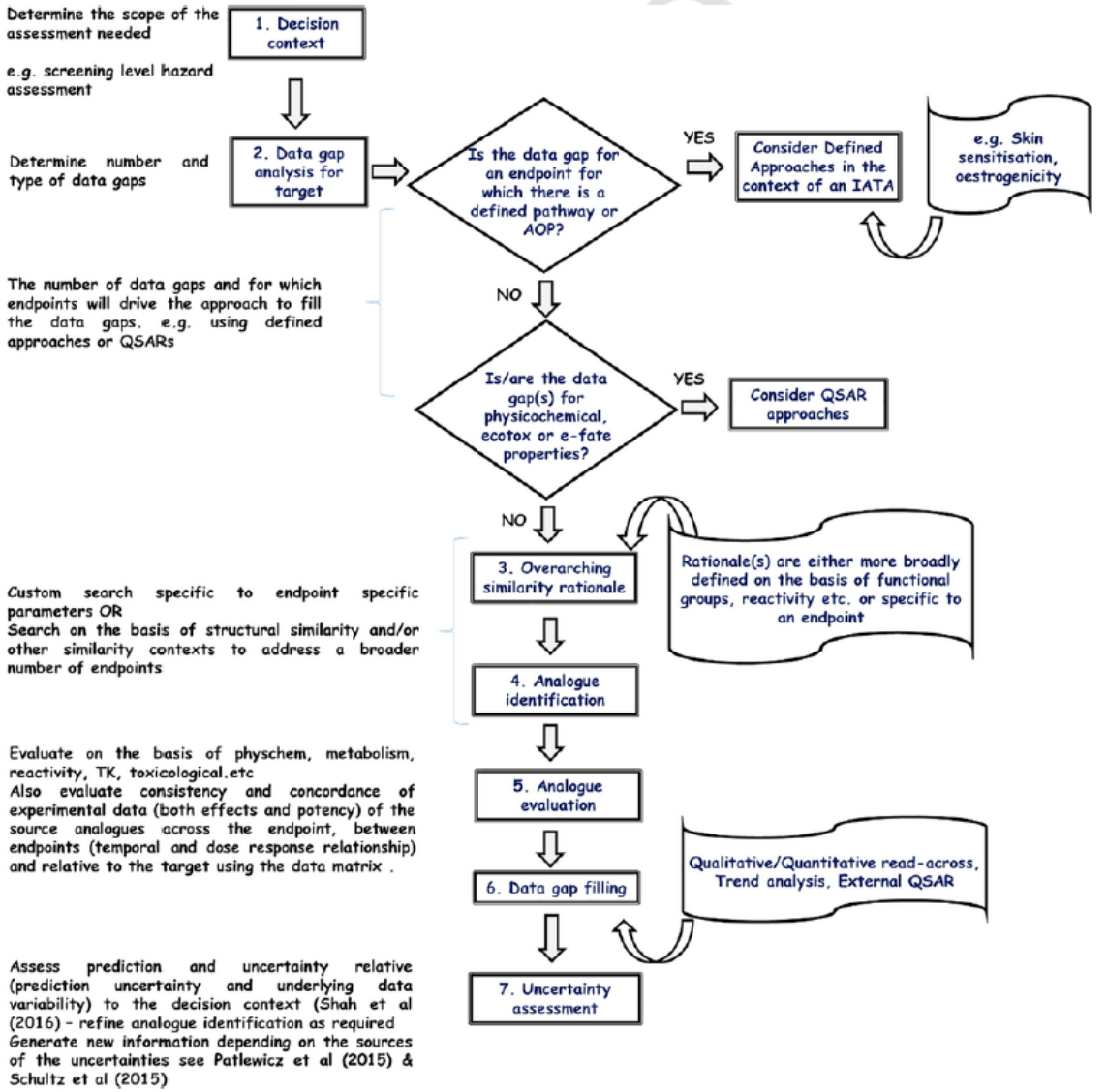


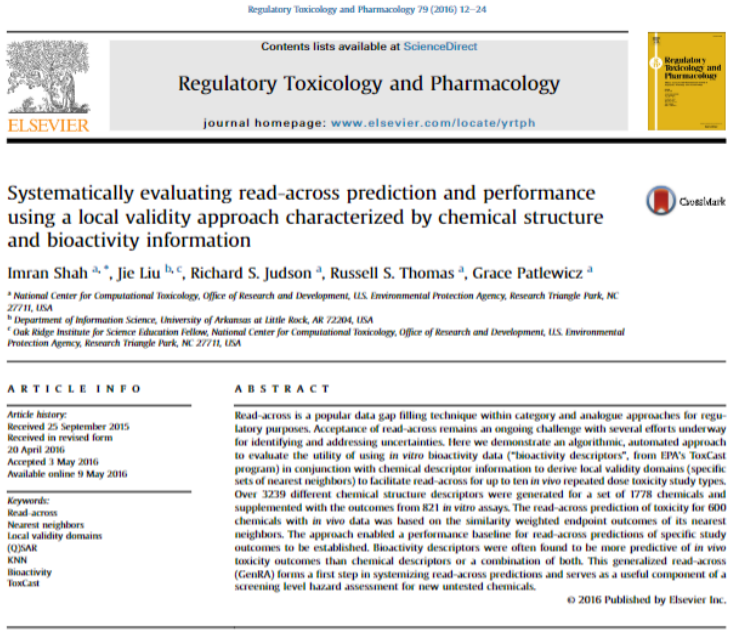
Fig. 9. A harmonised hybrid development and assessment framework.

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:

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

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

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

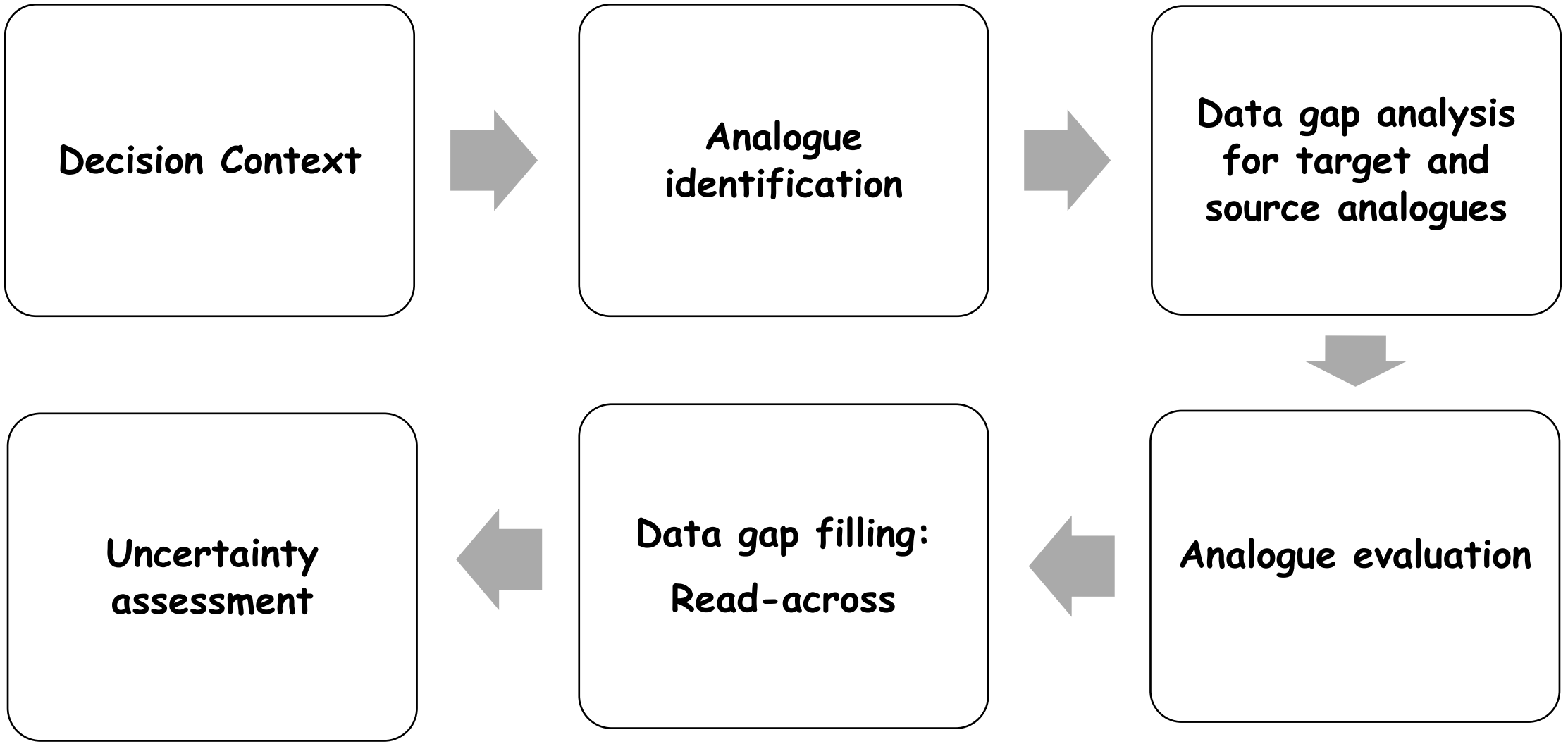
y_i = *predicted activity of chemical* (c_i)

x_j^{β} = *activity of* c_j in β

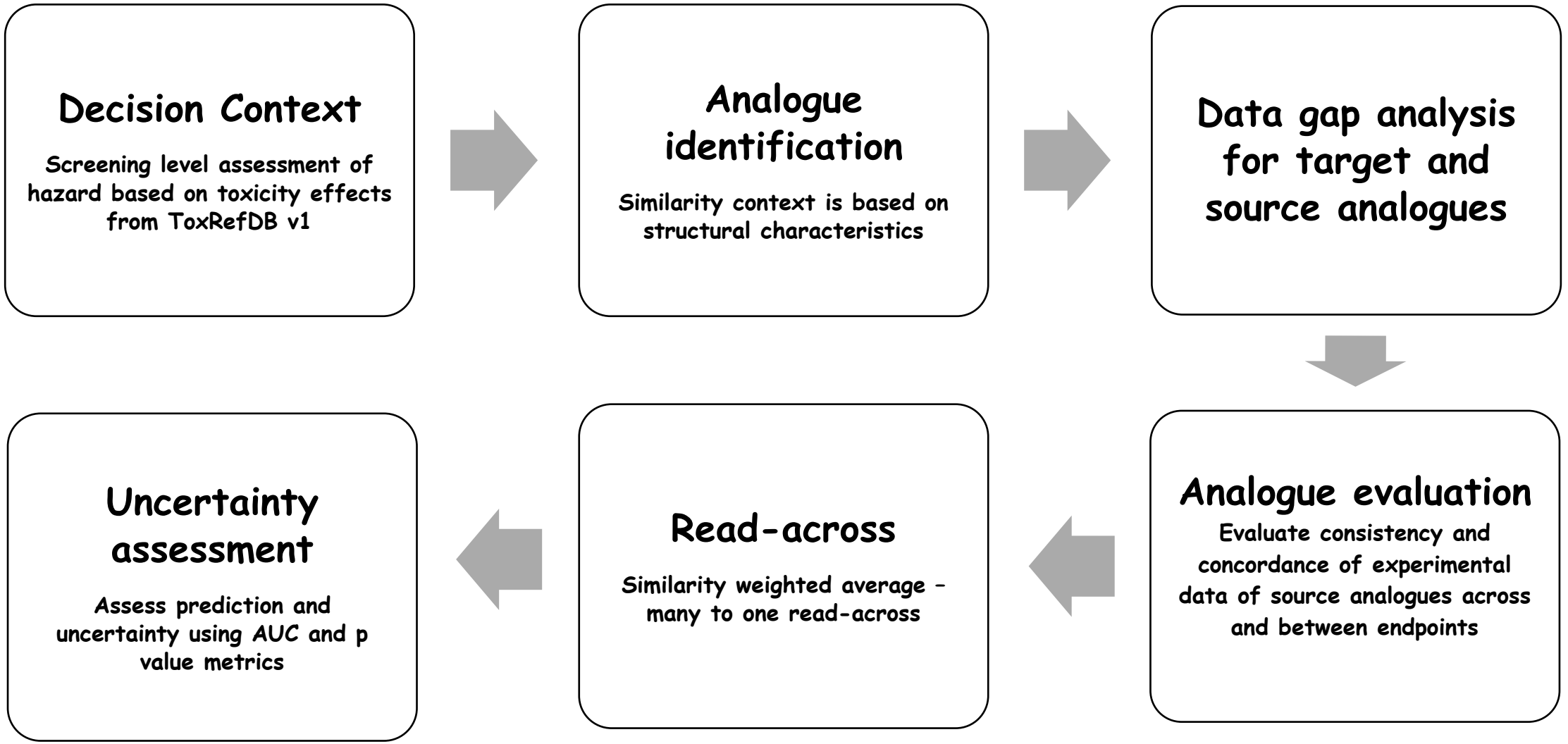
s_{ij}^{α} = *Jacccard similarity between* $\boldsymbol{x}_i^{\alpha}, \boldsymbol{x}_j^{\alpha}$

k = *up to* k *nearest neighbours*

Read-across workflow



Read-across workflow in GenRA v1.0



GenRA tool in reality

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

GenRA tools

- GenRA v3.0 webapp released February 2022

- 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)
- See https://github.com/patlewig/UNC_Rax for example to test out the tool with a specific acute toxicity example.

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.

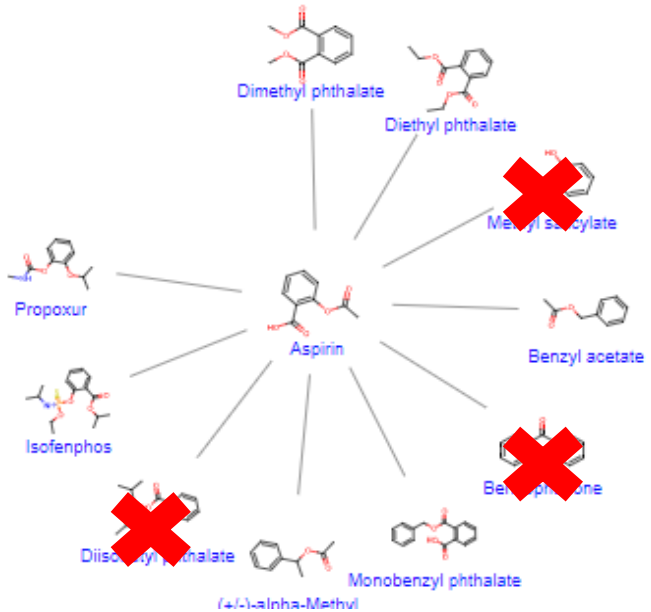
Contact: shah.imran@epa.gov

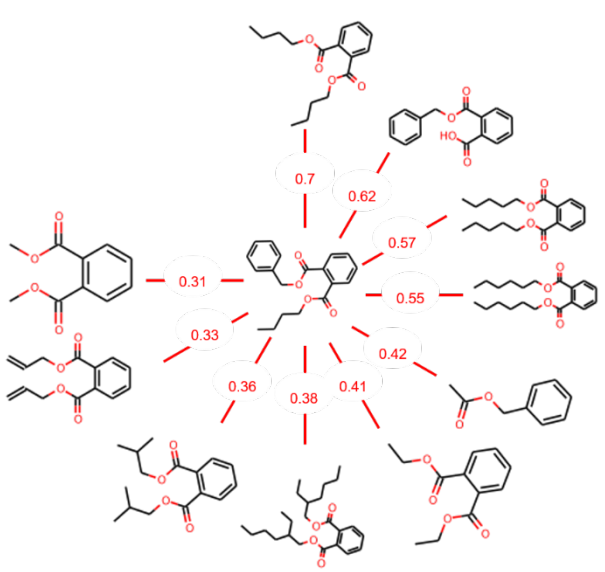
GenRA – Past and ongoing research

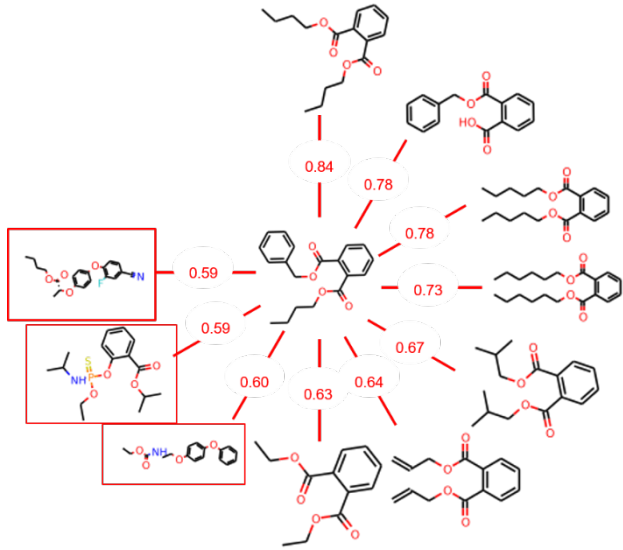
- Consideration of other information to define and refine the analogue selection & evaluation
  - physicochemical similarity (Helman et al 2018)
  - metabolic similarity (Boyce et al, 2022; Hagan et al, in prep; Groff et al, *in prep*)
  - reactivity similarity (Nelms et al 2018)
  - transcriptomics similarity (Tate et al, 2021)
- Transitioning to quantitative predictions of toxicity
  - Using GenRA to predict Lowest Observed Adverse Effect Level (LOAEL), acute oral (median lethal dose) 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 (*in prep*)

Physicochemical similarity

- Find structurally similar analogues and filter based on physicochemical considerations (filter out) vs find similar analogues with respect to structure and physicochemical similarity at the same time (Search expansion).







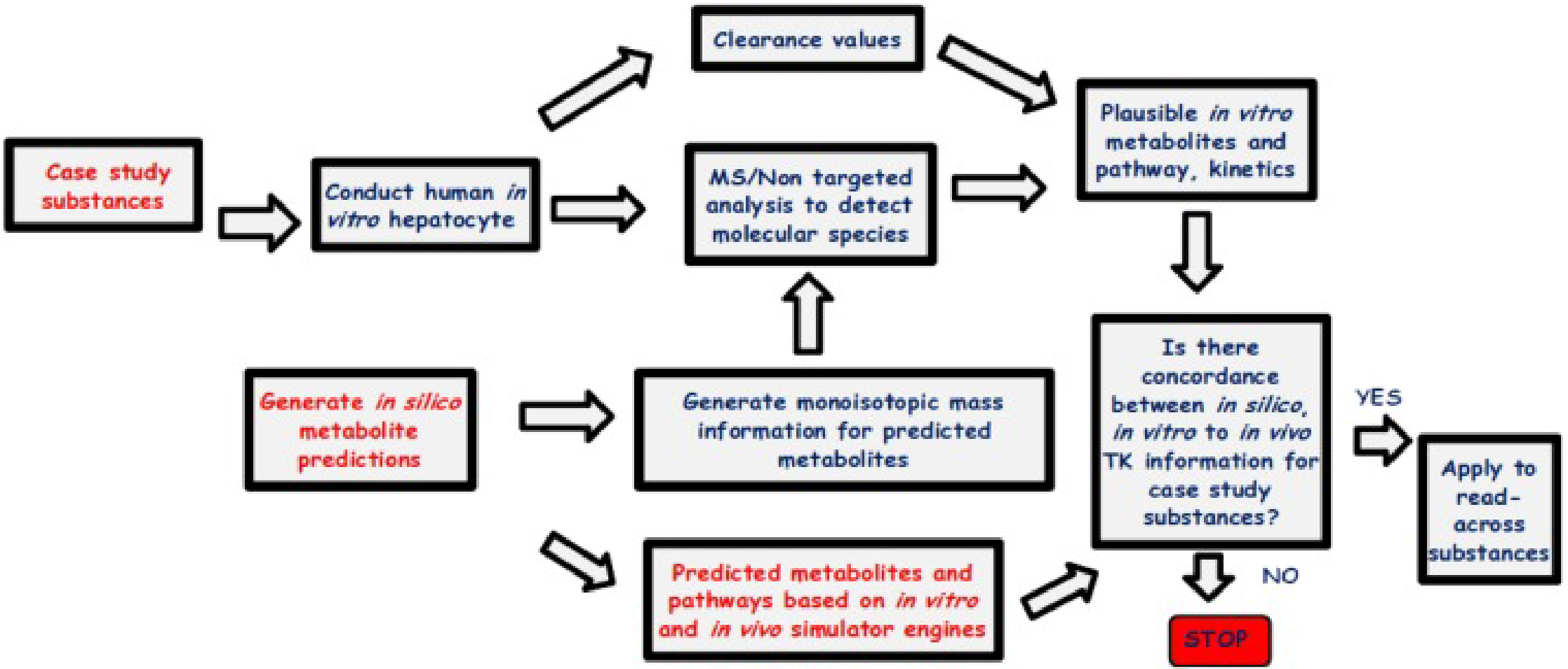
- Similarity search using Jaccard distance of Morgan chemical fingerprints to find source analogues. (Default of 10 nearest neighbors (k=10))
- Calculate physchem similarity between target and source analogues using a generalised Jaccard similarity metric
- Reduce neighborhood based on the physchem similarity threshold

- Similarity search using weighted sum of Morgan chemical fingerprint distance and physchem distance to find source analogues

Metabolic similarity

- Metabolic similarity is a key consideration in read-across but approaches to quantify the contribution metabolism plays are still evolving.
- Characterising metabolic similarity could encompass codifying the structural similarity of the metabolites, the similarity in metabolite profile, the sequence of transformations as well as the transformations themselves.
- But availability of metabolite information is limited – how does reported metabolic information relate to predicted metabolic information from different tools…
- Several avenues being explored through proof-of-concept studies.
  - 1) Generating in vitro data in primary hepatocytes and using Mass Spec to identify the metabolites produced relative to a suspect screening list derived from running different expert systems to predict metabolism
  - 2) Evaluate the performance of the expert systems relative to reported literature data
  - 3) Explore ways of codifying metabolic similarity

Proof of concept workflow for (1)



Boyce et al, under review

Selected expert systems evaluated in (2)

Model	Availability	Module/Species	Prediction Settings	Number of Predicted Metabolites
BioTransformer	Free (http://biotransformer.ca)	Human	Phase I: 2 steps Phase II: 1 step	3464
Meteor	Commercial (https://www.lhasalimited.org/)	Mammal*	Default	714
Toolbox	Free (https://qsartoolbox.org/)	Rat (S9, *in vitro*), Rat (*in vivo*)	Default	194 (*in vitro*), 236 (*in vivo*)
TIMES[‡]	Commercial (http://oasis-lmc.org)	Rat (S9, *in vitro*), Rat (*in vivo*)	Default	283 (*in vitro*), 570 (*in vivo*)
SyGMa	Free (https://sygma.readthedocs.io)	Human	Phase I: 2 steps Phase II: 1 step	5215

Encoding metabolic similarity (3)

	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
			0.0	0.5	0.0
			1.0	0.0	0.0
			0.0	1.0	0.0
DTXSID00379884	0.0	0.0	0.0	0.0	1.0

Creating custom fingerprints to characterise metabolic transformations
Boyce et al (2022)

Encoding metabolic similarity (3)

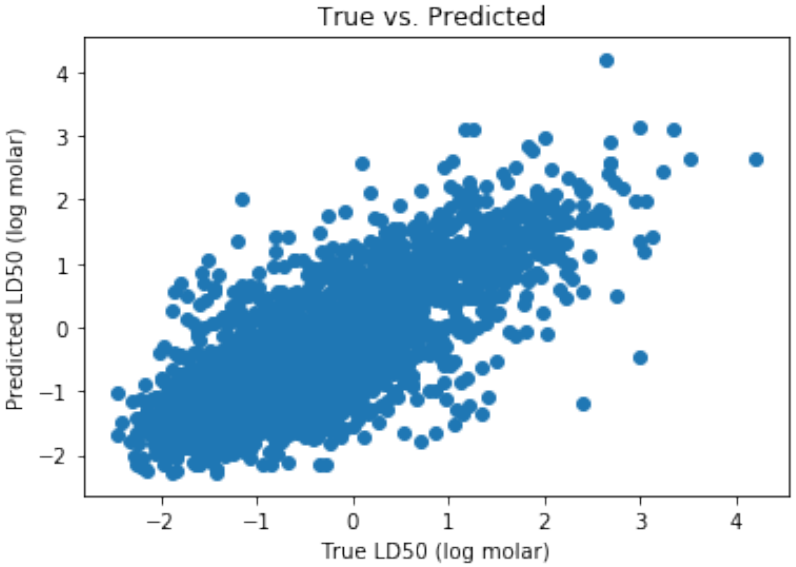
- The structure of transformation pathways naturally lend themselves to graph representations, for which a number of different methods can be applied, including graph kernels, to determine the pairwise similarity.
- Currently exploring the correlation between metabolic vs structural similarities on neighbourhood formation for a set of 18 POC substances using predicted metabolites that have been generated using BioTransformer (CYP450 mode), and TIssue MEtabolism Simulator (TIMES) (*in vivo* and *in vitro*).
- Metabolic similarity is quantified by the graph kernel, Weisfeiler-Lehman (WL)

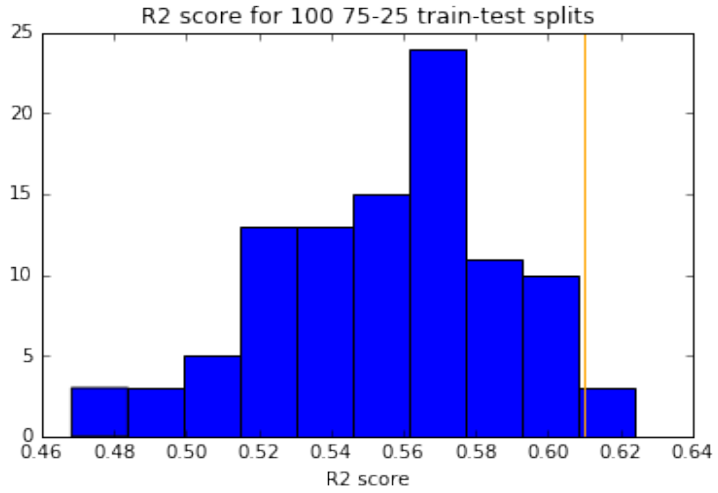
Encoding metabolic similarity (3)

- Graph kernels are functions that can measure similarity between graphs by operating directly on the graph structures. Two graphs (metabolic trees) could be considered similar if they are composed of nodes (substances) with similar neighbourhoods
- The WL kernel creates a feature vector of counts of iteratively generated graph labels that are constructed by creating a superset of a node's neighbours and then hashing that superset into a new, condensed label that contains both structural and contextual information
- Our similarity metric is then defined by the Jaccard similarity between the feature vectors generated by the kernel function

Quantitative predictions: Acute toxicity

- 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



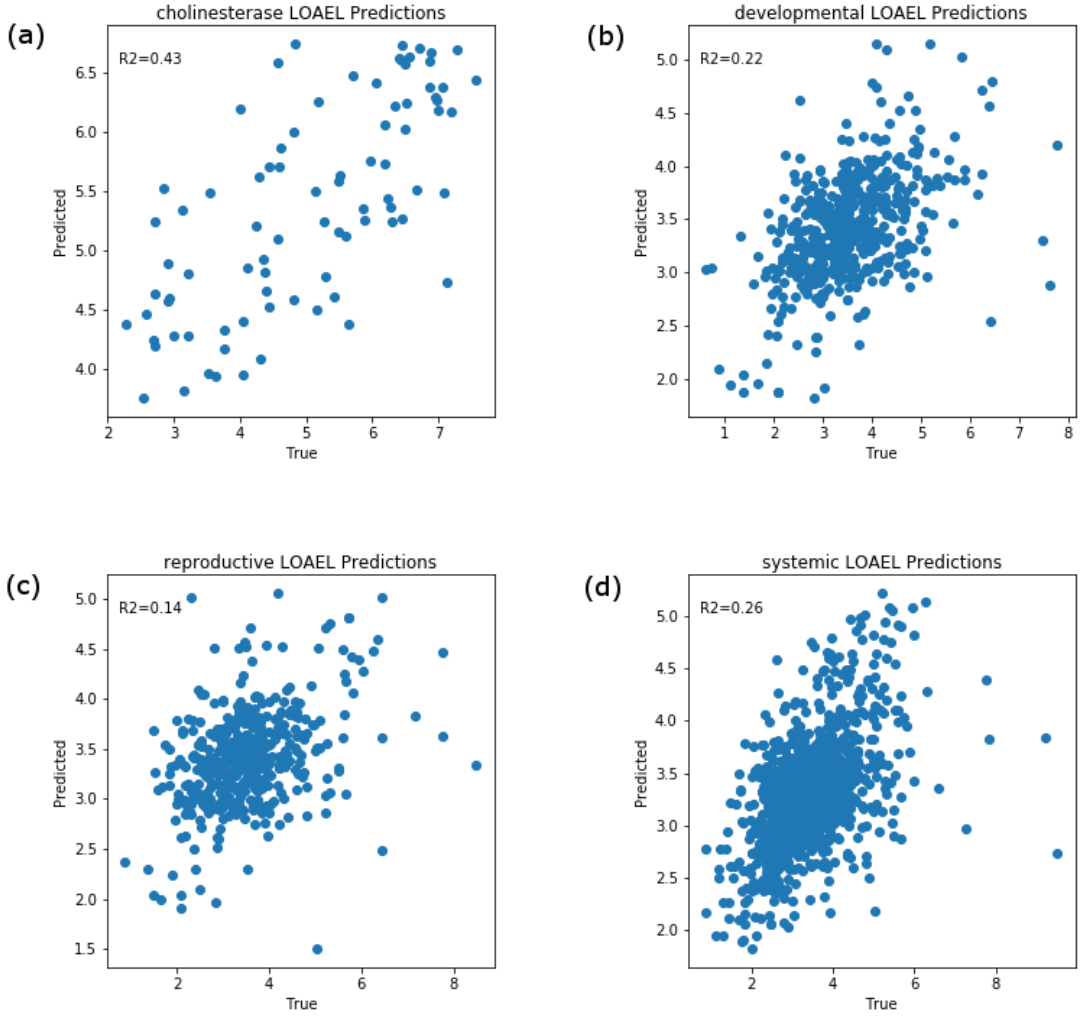


- Linear regression used to fit predicted and observed LD50 values
- R^2 = 0.61
- RMSE = 0.58

Helman et al., 2019a

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

Quantitative predictions: LOAEL values



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

GenRA Version 2 Highlights

- Version 2:
- Maintained existing read-across workflow
- Complete rebuild of GenRA Version 1.0
- ToxRefDB updated from Version 1 to Version 2
- ToxCast data updated
- Chemical fingerprints recomputed to factor in additional substances in the DSSTox database that had been registered since initial release
- Ability to search for analogues without prefiltering on the basis of ToxRefDB data
- GenRA decoupled from the Dashboard i.e. an independent application but one which is still linked to the Dashboard

EPA United States Environmental Protection Agency

GenRA Version 3 Highlights

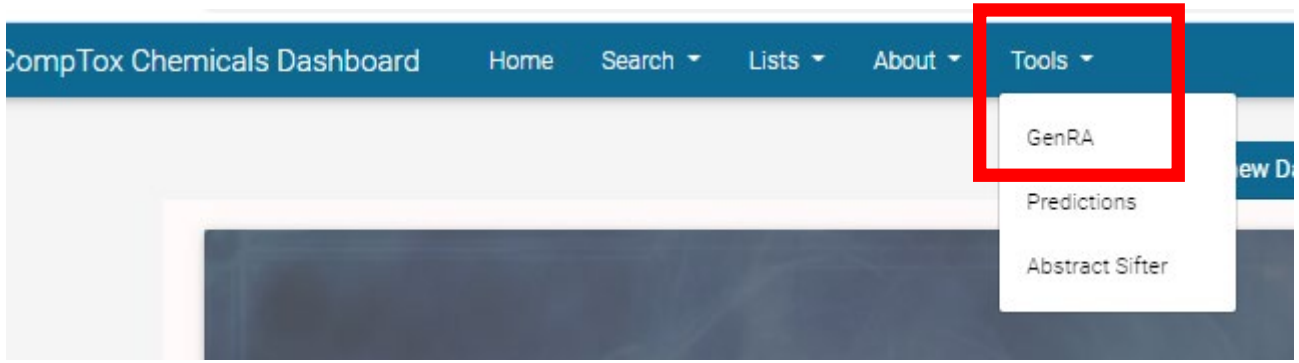
- Version 3
- UI rebuilt using AG Grid to provide more out of the box interactivity
- Custom fingerprints (users can specify fingerprint combinations based on existing fingerprints provided)
- Ketcher drawing palette to allow SMILES/MOL to be introduced and predictions to be made for substances not already within the Dashboard
- Contact email added to track bugs/refinements

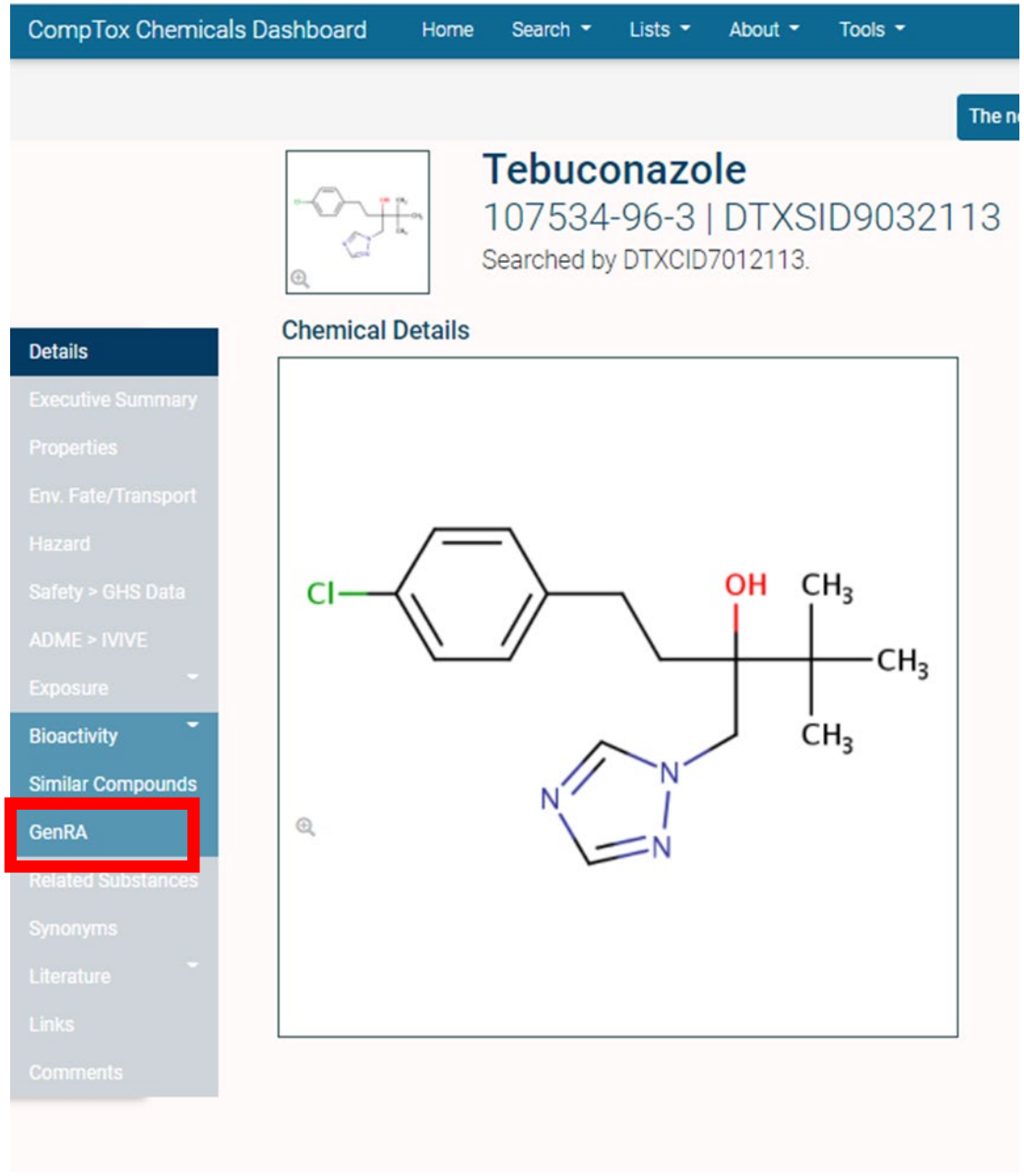
GenRA Version 3

Main entry point is from the portal comptox.epa.gov

However, can be accessed from the landing page within the Dashboard for a specific chemical or from the Tools menu within the Dashboard

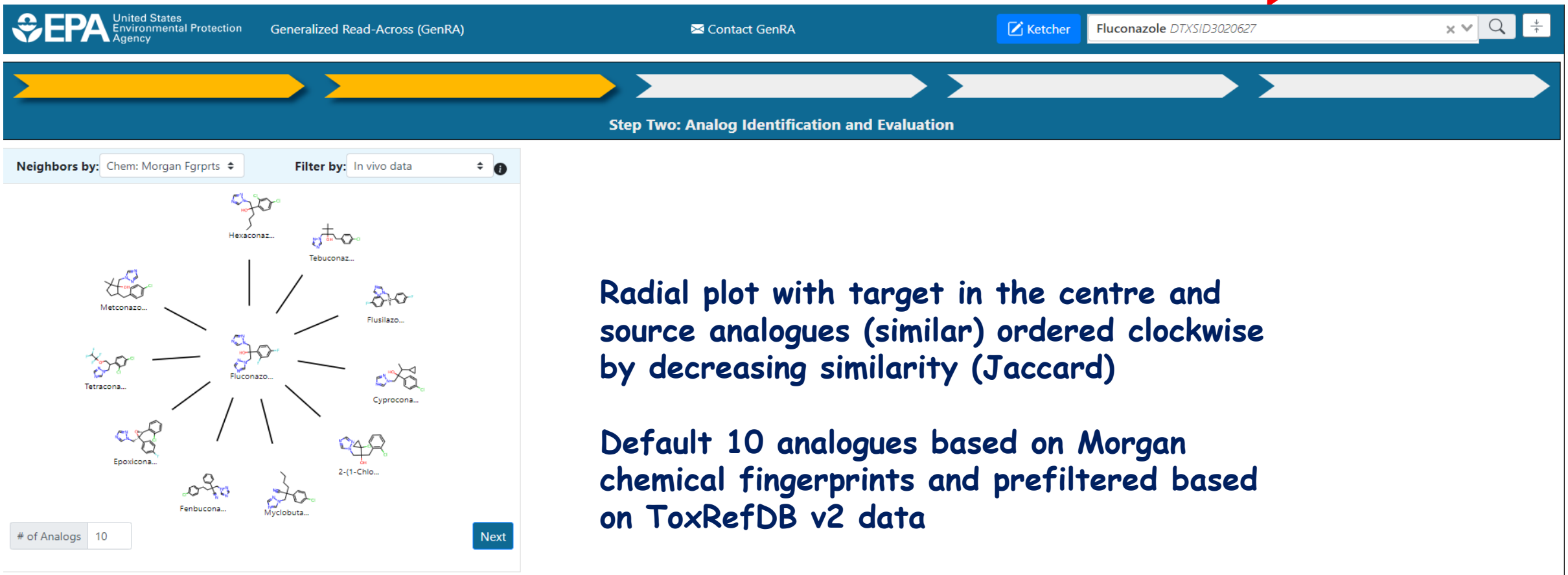
Alternative entry points





GenRA v3 in practice

Search for a chemical of interest (target) using the search box

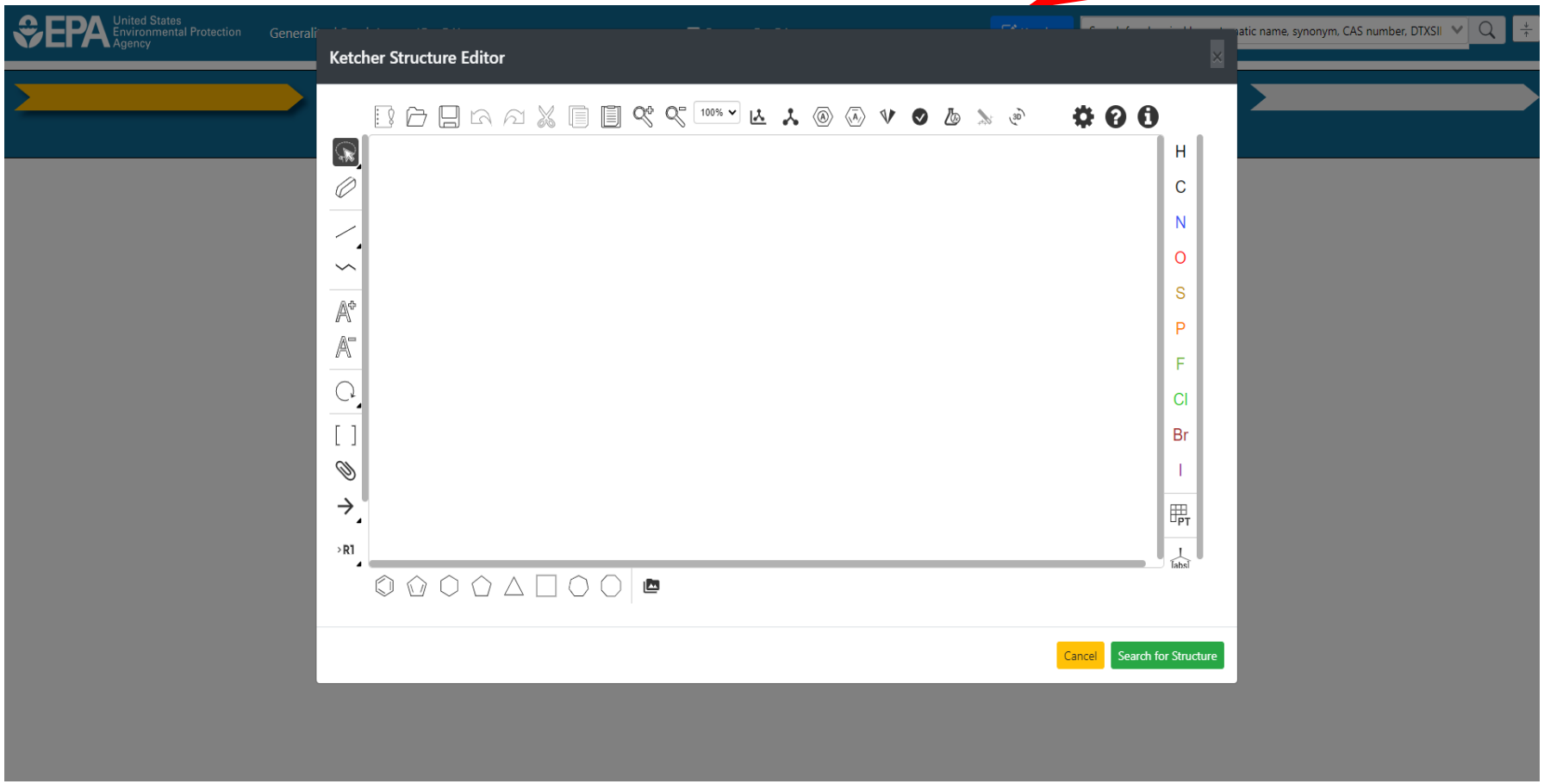


Radial plot with target in the centre and source analogues (similar) ordered clockwise by decreasing similarity (Jaccard)

Default 10 analogues based on Morgan chemical fingerprints and prefiltered based on ToxRefDB v2 data

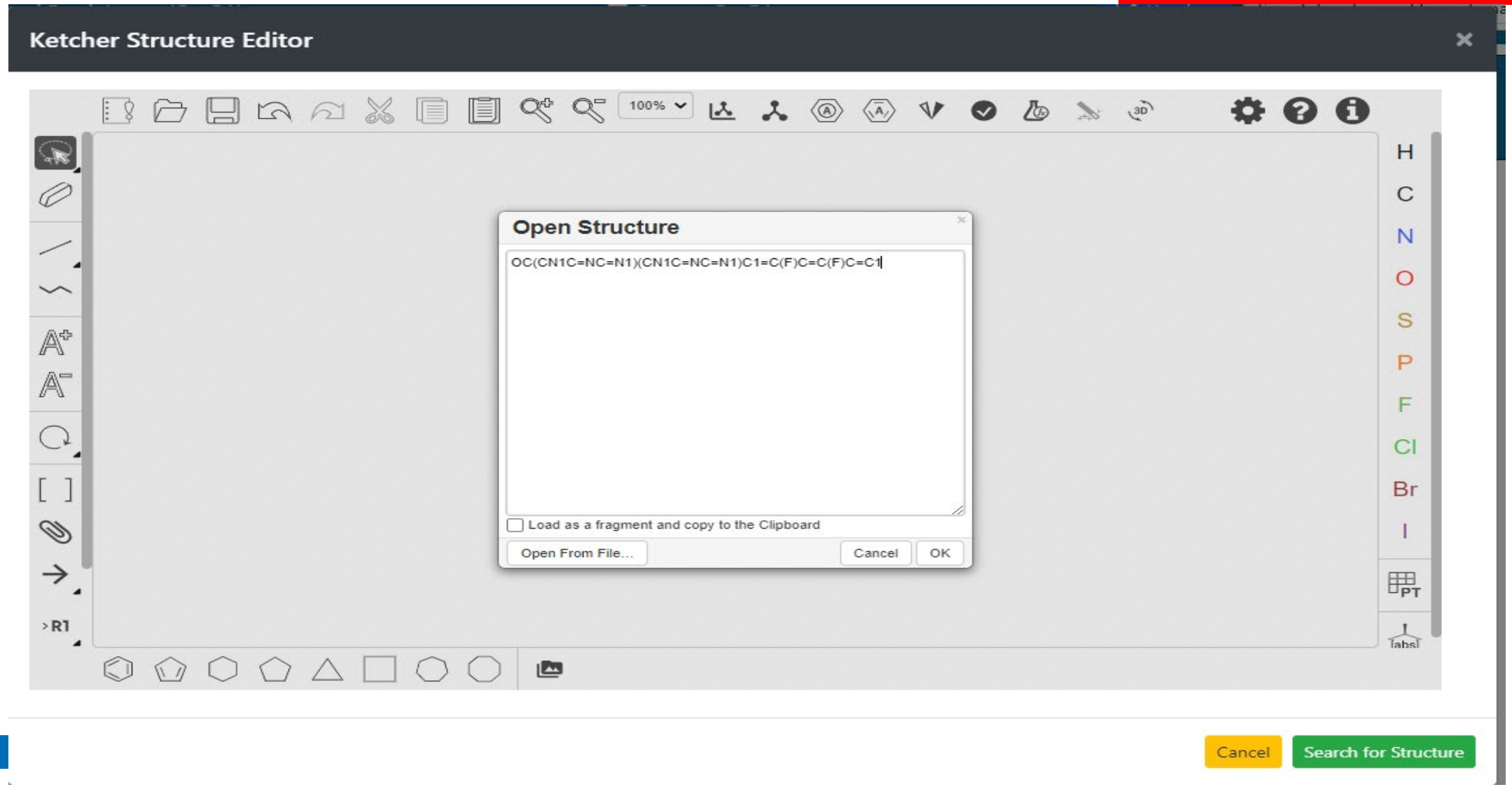
GenRA v3 in practice

Search for a chemical of interest (target) using the Ketcher

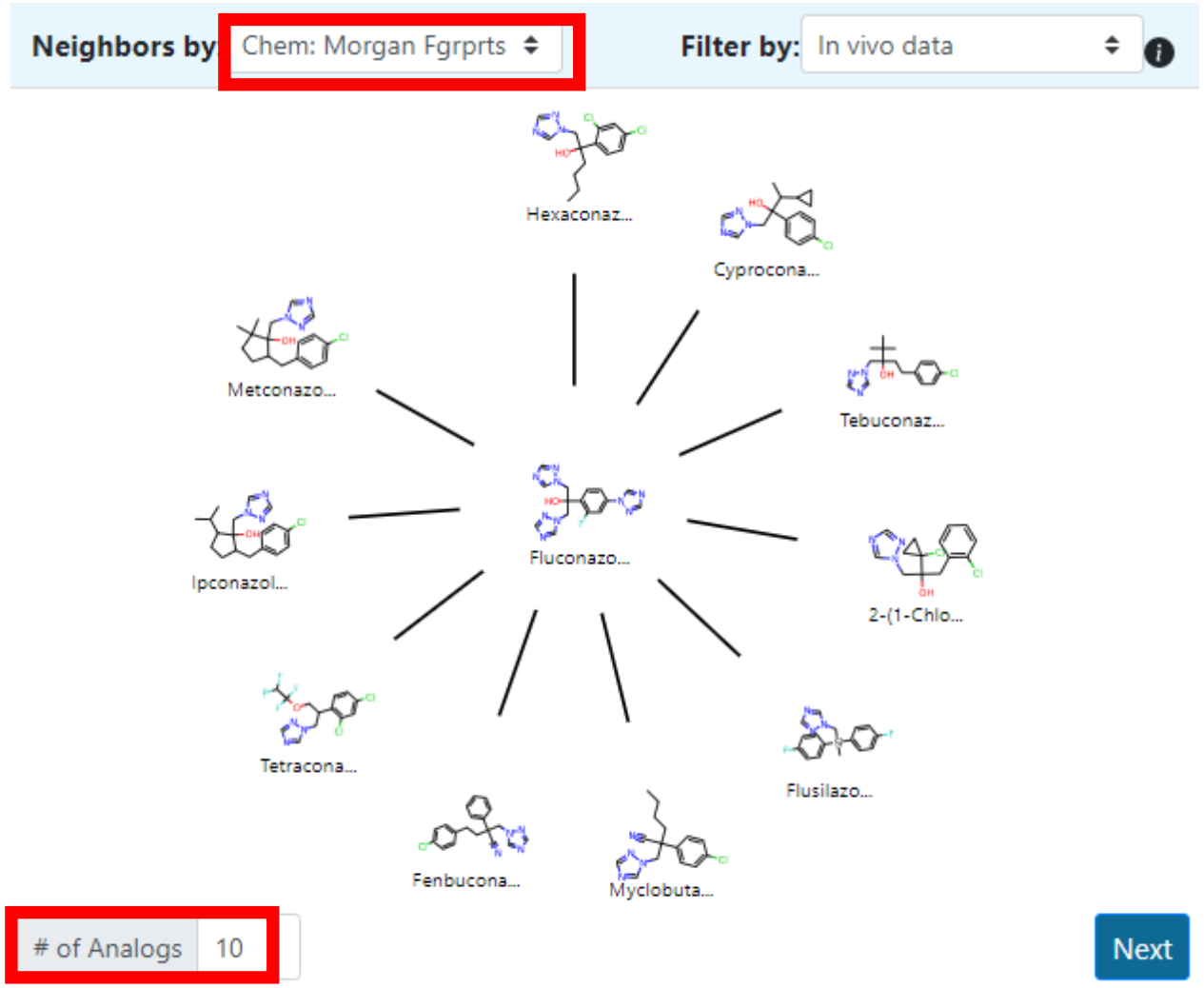


GenRA v3 in practice

Search for a chemical of interest (target) using the Ketcher



GenRA v3 in practice

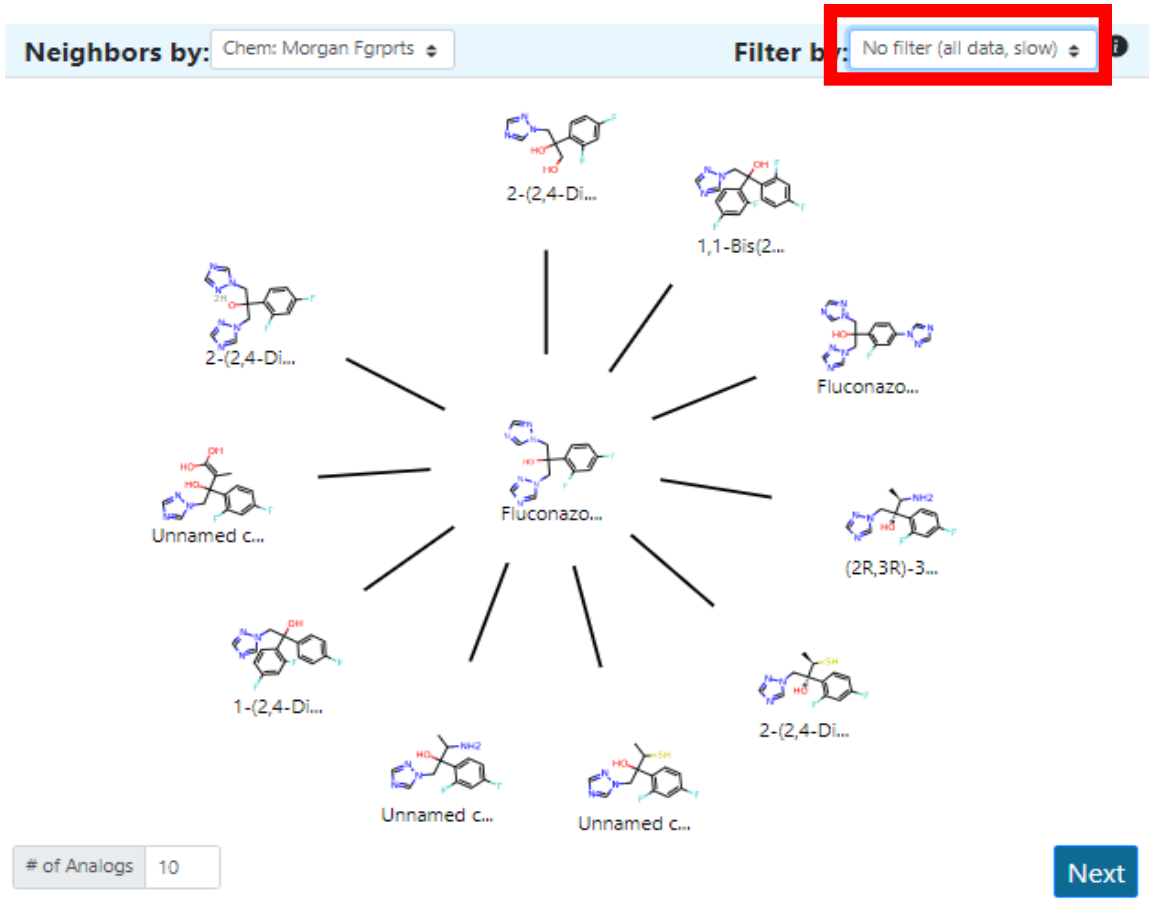


Radial plot with target in the centre and source analogues (similar) ordered clockwise by decreasing similarity (Jaccard)

Default 10 analogues based on Morgan chemical fingerprints and prefiltered based on ToxRefDB v2 data

Can update to change what features are used to characterise substances and the number of analogues returned

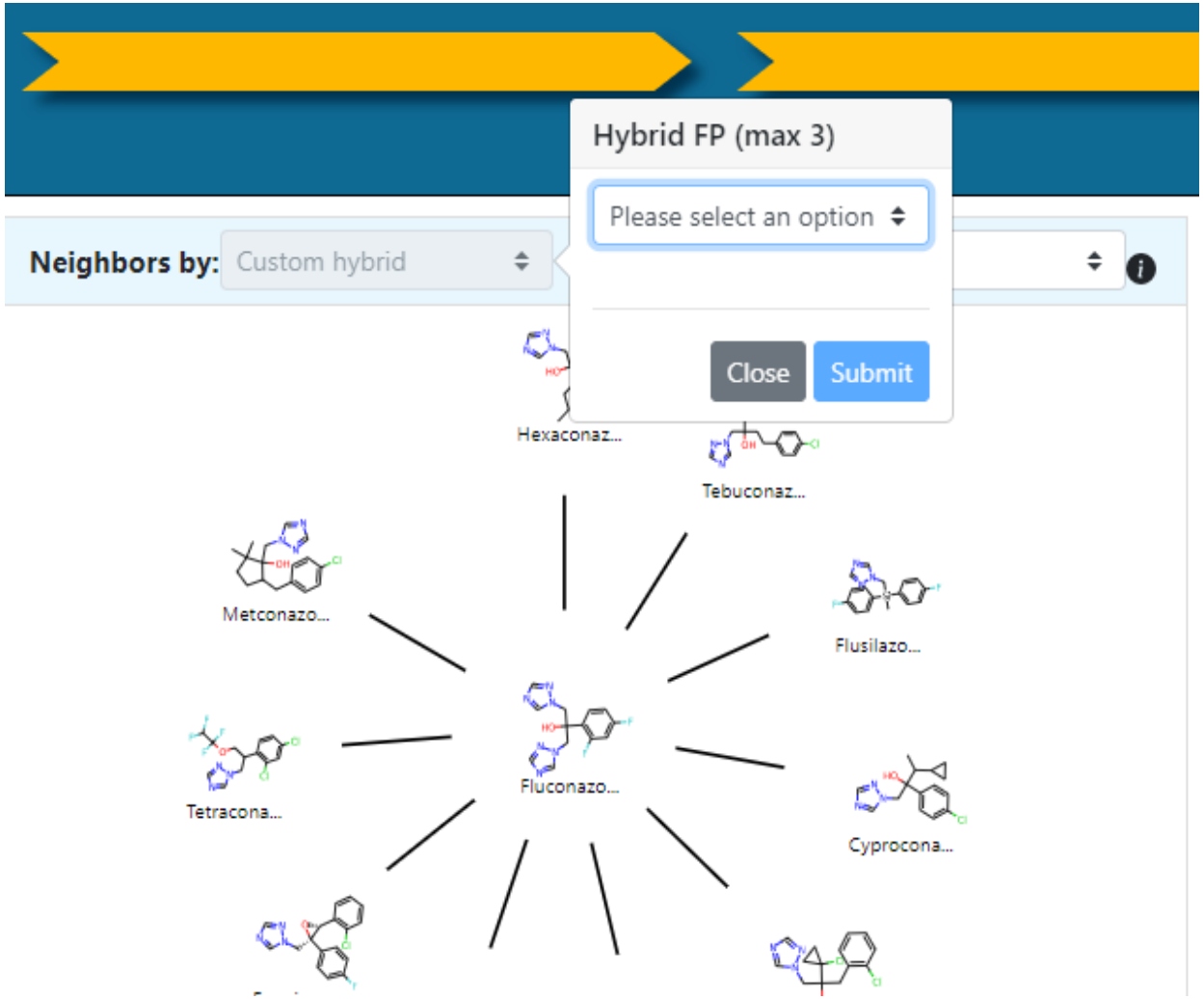
GenRA v3 in practice



Update radial plot to return analogues irrespective of ToxRefDB v2 data

Caution! This can be quite slow

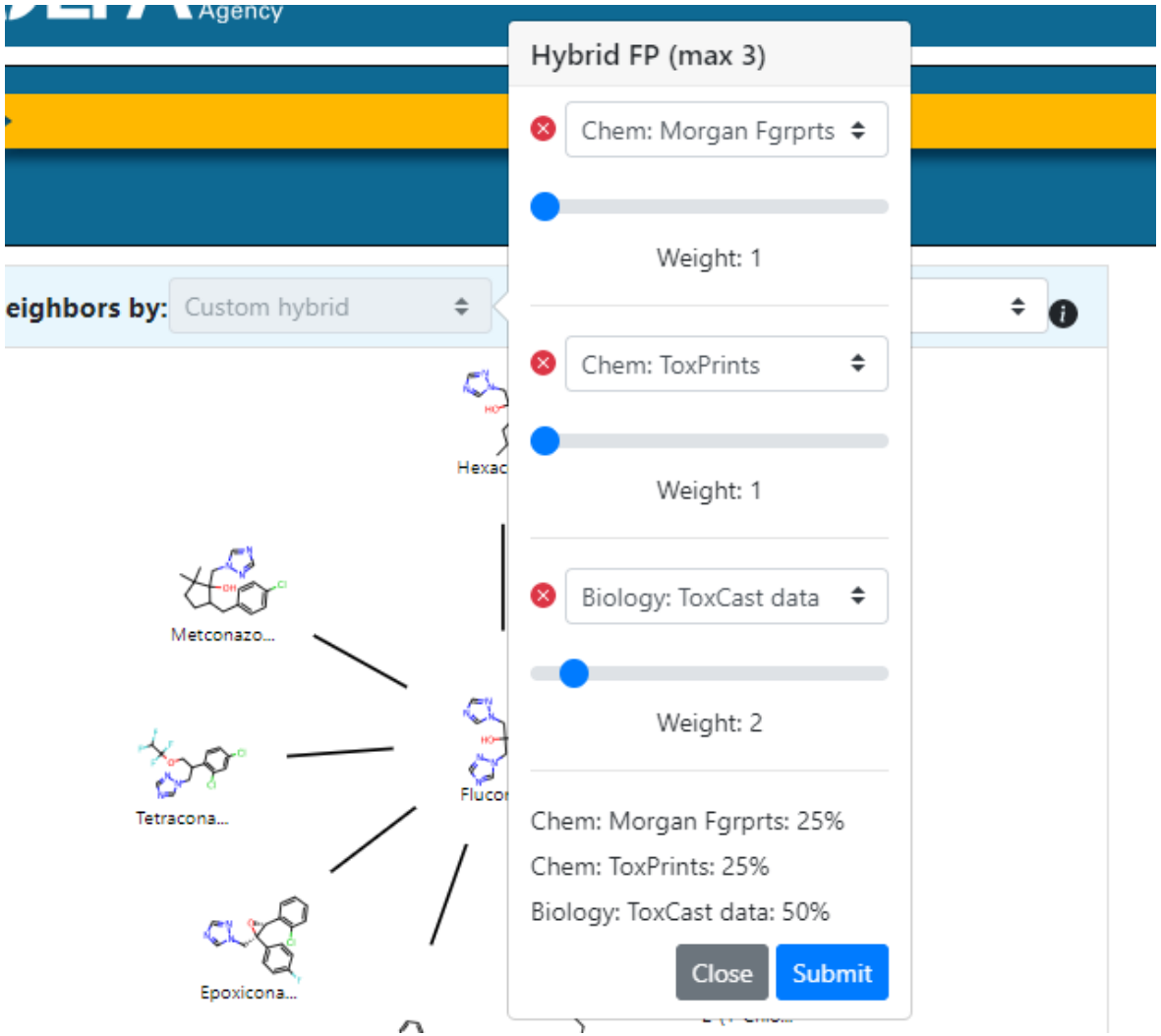
GenRA v3 in practice



Custom Fingerprints

Choose up to 3 fingerprints e.g. 50% ToxCast vs 50% Chemical Morgan Fingerprints & 25% ToxPrints

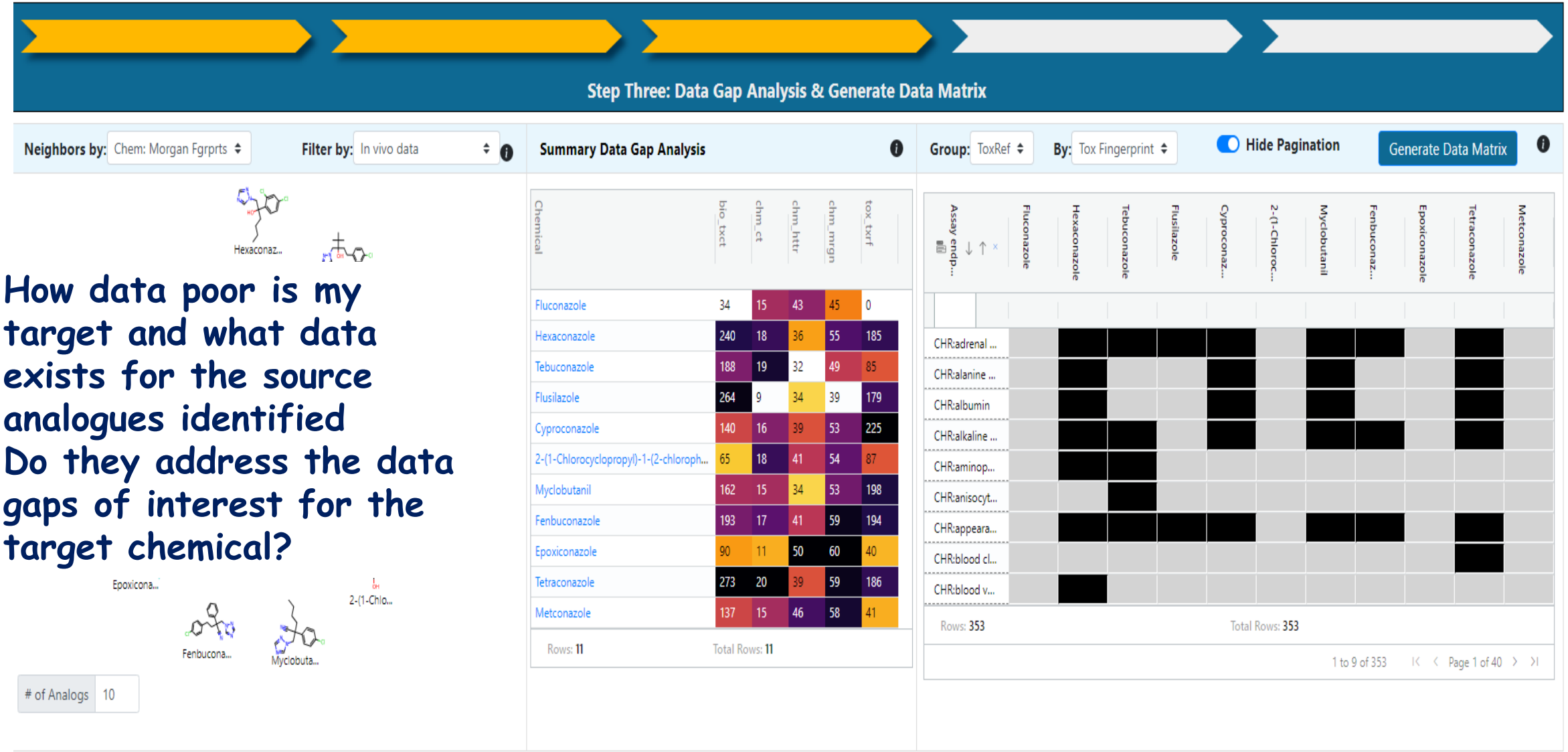
GenRA v3 in practice



Custom Fingerprints

Choose up to 3 fingerprints
e.g. 50% ToxCast vs 25% Chemical Morgan Fingerprints & 25% ToxPrints

GenRA v3 in practice

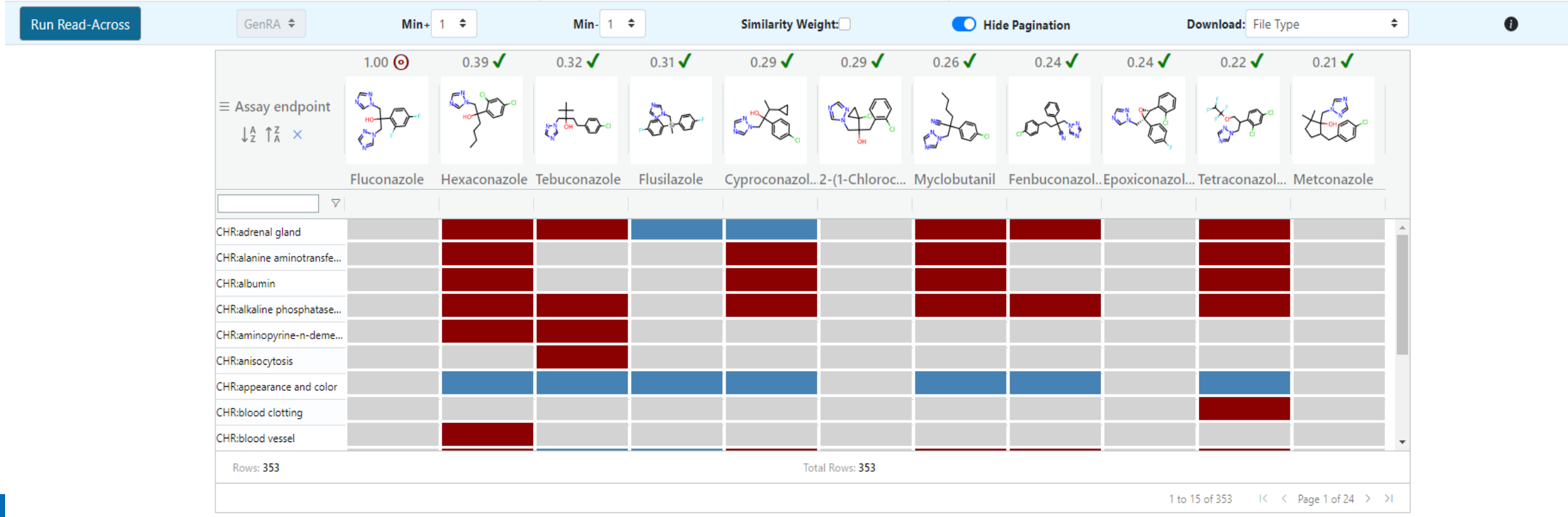


- How data poor is my target and what data exists for the source analogues identified
- Do they address the data gaps of interest for the target chemical?

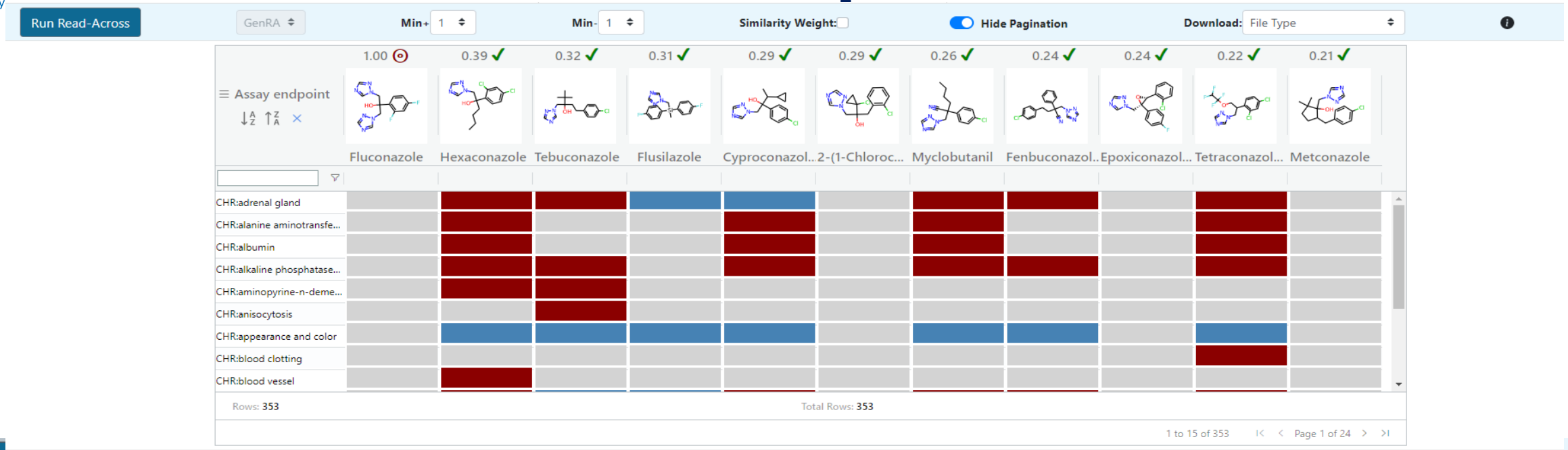
GenRA v3 tool in practice

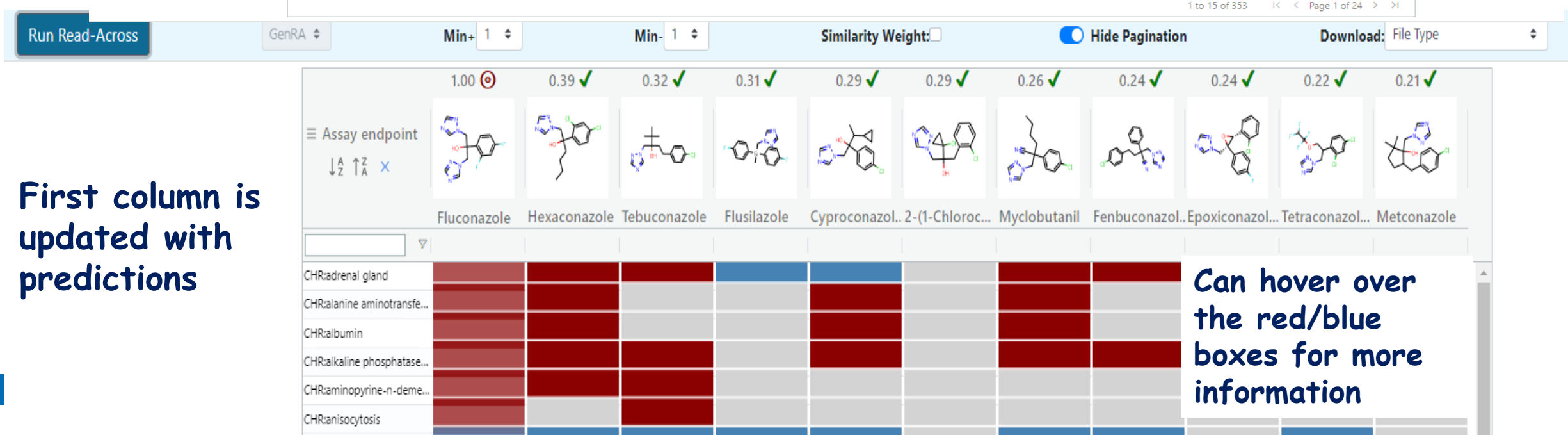
What is the consistency and concordance across my source analogues?
Should I deselect analogues from consideration from the entire set of predictions?
Should I consider subcategorising the analogues selected?

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



GenRA v3 in practice





First column is updated with predictions

Can hover over the red/blue boxes for more information

EPA United States Environmental Protection Agency

GenRA v3 in practice

- Database underpinning GenRA v3.0: ToxRefDB v2
  - Different study types and effects within them are predicted e.g. chronic_liver is annotated as CHR_liver
  - Negative effects are inferred from guideline profiles which define the required tests for each study type. The assumption is that the study required an evaluation but no effects were reported.
  - Positive results – min dose at which toxicity effects are observed in a study
- Prediction: Similarity weighted activity
- Performance is categorised by the Area under the Curve (AUC) of the Receiving Operating Characteristic (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 v3 in practice

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

chem_id	DTXCID10627	DTXCID10627_uni	DTXCID2014653	DTXCID2014653_u	DTXCID7012113	DTXCID7012113_u	DTXCID704235	DTXCID704235_ur	DTXCID8012601	DTXCID8012601_u	DTXCID1024338	DTXCID1024338_u	DTXCID304315	DTXCID304315_ur	D
role	target		analog		analog		analog		analog		analog		analog		a
preferred name	Fluconazole		Hexaconazole		Tebuconazole		Flusilazole		Cyproconazole		2-(1-Chlorocyclopropyl)-1-(2-chlorc		Myclobutanil		F
dsstox_sid	DTXSID3020627		DTXSID4034653		DTXSID9032113		DTXSID3024235		DTXSID0032601		DTXSID3044338		DTXSID8024315		D
dsstox_cid	DTXCID10627		DTXCID2014653		DTXCID7012113		DTXCID704235		DTXCID8012601		DTXCID1024338		DTXCID304315		D
molecular weight	306.277		314.21		307.82		315.399		291.78		312.19		288.78		
similarity	1		0.389		0.324		0.312		0.289		0.286		0.256		
CHR:adrenal gland	GenRA Pos; ACT=None; AUC=0.0; pval=0.945		4.7	mg/kg/day	4.39	mg/kg/day	no_effect		no_effect		no_data		393.5	mg/kg/day	
CHR:alanine aminotransferas	GenRA Pos; ACT=None; AUC=0; pval=1		50	mg/kg/day	no_data		no_data		12.1	mg/kg/day	no_data		40	mg/kg/day	r
CHR:albumin	GenRA Pos; ACT=None; AUC=0; pval=1		50	mg/kg/day	no_data		no_data		12.1	mg/kg/day	no_data		40	mg/kg/day	r
CHR:alkaline phosphatase (al	GenRA Pos; ACT=None; AUC=0; pval=1		10	mg/kg/day	482.9	mg/kg/day	no_data		3.2	mg/kg/day	no_data		15.68	mg/kg/day	
CHR:aminopyrine-n-demethy	GenRA Pos; ACT=None; AUC=0; pval=1		4.7	mg/kg/day	482.9	mg/kg/day	no_data		no_data		no_data		no_data		r
CHR:anisocytosis	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		77.3	mg/kg/day	no_data		no_data		no_data		no_data		r
CHR:appearance and color	GenRA Neg; ACT=None; AUC=0; pval=1		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:blood clotting	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		no_data		no_data		no_data		r
CHR:blood vessel	GenRA Pos; ACT=None; AUC=0; pval=1		50	mg/kg/day	no_data		no_data		no_data		no_data		no_data		r
CHR:body weight	GenRA Pos; ACT=None; AUC=0.0; pval=0.94		6.1	mg/kg/day	no_effect		no_effect		3.2	mg/kg/day	no_data		39.21	mg/kg/day	
CHR:bone	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		no_data		no_data		no_data		r
CHR:bone marrow	GenRA Neg; ACT=None; AUC=0; pval=1		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:brain	GenRA Neg; ACT=None; AUC=0.0; pval=0.93		no_effect		no_effect		no_effect		no_effect		no_data		393.5	mg/kg/day	r
CHR:calcium	GenRA Pos; ACT=None; AUC=0; pval=1		50	mg/kg/day	no_data		no_data		12.1	mg/kg/day	no_data		no_data		r
CHR:cholesterol	GenRA Pos; ACT=None; AUC=0; pval=1		50	mg/kg/day	no_data		13	mg/kg/day	12.1	mg/kg/day	no_data		no_data		
CHR:clinical signs	GenRA Pos; ACT=None; AUC=0.0; pval=0.965		no_effect		no_effect		27	mg/kg/day	3.2	mg/kg/day	no_data		no_effect		
CHR:cytochrome p450, nos	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		3.2	mg/kg/day	no_data		no_data		r
CHR:ear	GenRA Neg; ACT=None; AUC=0; pval=1		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:epididymis	GenRA Neg; ACT=None; AUC=0.0; pval=0.93		no_effect		no_effect		no_effect		13.17	mg/kg/day	no_data		125	mg/kg/day	r
CHR:erythrocyte (rbc) count d	GenRA Neg; ACT=None; AUC=0.0; pval=0.84		no_effect		no_data		no_data		no_effect		no_data		40	mg/kg/day	r
CHR:esophagus	GenRA Neg; ACT=None; AUC=0; pval=1		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:eye	GenRA Neg; ACT=None; AUC=0.0; pval=0.86		no_effect		77.3	mg/kg/day	no_effect		no_effect		no_data		no_effect		r
CHR:food consumption	GenRA Neg; ACT=None; AUC=0.0; pval=0.99		267	mg/kg/day	no_effect		no_effect		12.1	mg/kg/day	no_data		40	mg/kg/day	r
CHR:full gross necropsy	GenRA Neg; ACT=None; AUC=0; pval=1		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:gamma glutamyl transfe	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		no_data		no_data		40	mg/kg/day	r
CHR:globulins	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		12.1	mg/kg/day	no_data		no_data		r
CHR:glucose	GenRA Neg; ACT=None; AUC=0.0; pval=0.875		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:glutathione-s-transferas	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		3.2	mg/kg/day	no_data		no_data		r
CHR:heart	GenRA Neg; ACT=None; AUC=0.0; pval=0.85		no_effect		no_effect		no_effect		no_effect		no_data		no_effect		r
CHR:hematocrit (hct)	GenRA Pos; ACT=None; AUC=0; pval=1		no_data		no_data		no_data		no_data		no_data		no_data		

EPA United States Environmental Protection Agency

GenRA v3 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.

EPA United States Environmental Protection Agency

GenRA v3.1 – released Sept 2022

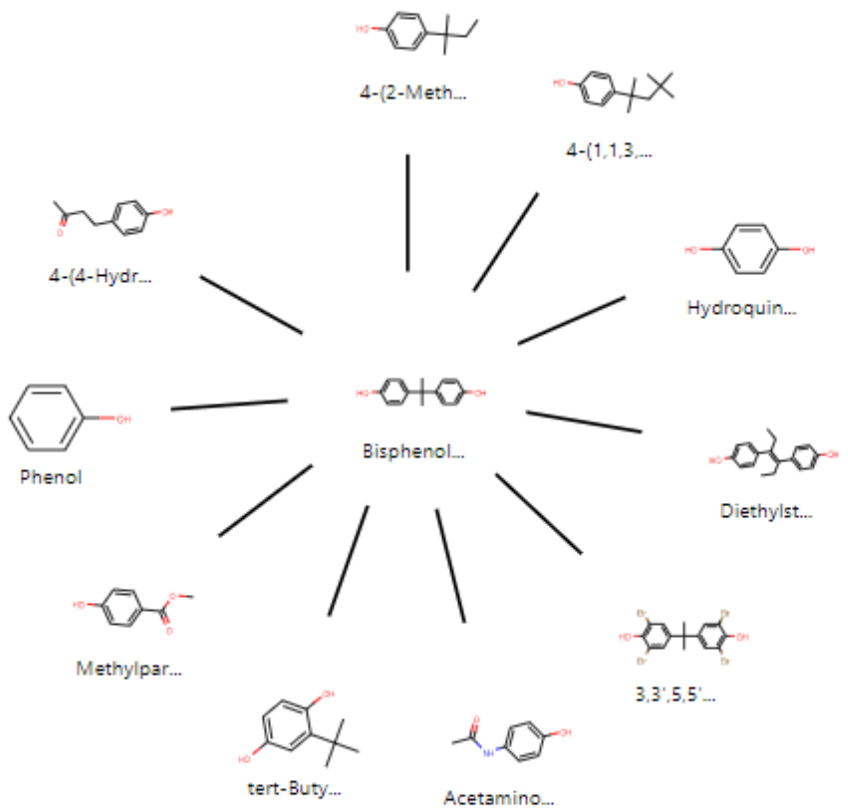
- Assessment of physchemicochemical similarity
- Network view
- New fingerprints that capture other NAM data
- Potency predictions using ToxRefDB
- Other data beyond in vivo toxicity endpoints

EPA United States Environmental Protection Agency

GenRA v3.1

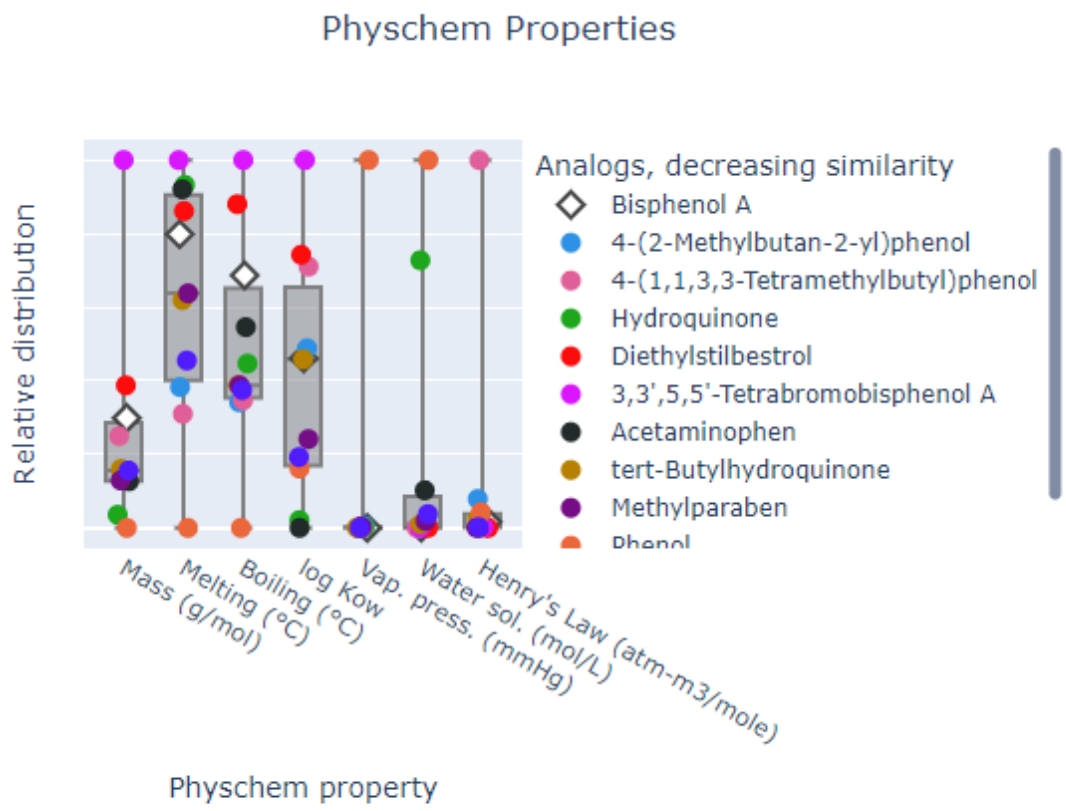
Step Two: Analog Identification and Evaluation

Neighbors by: Chem: Morgan Fgrprts **Filter by:** ToxRef data



# of Analogs 10 | Physchem Data | Neighborhood Exploration | Loading...

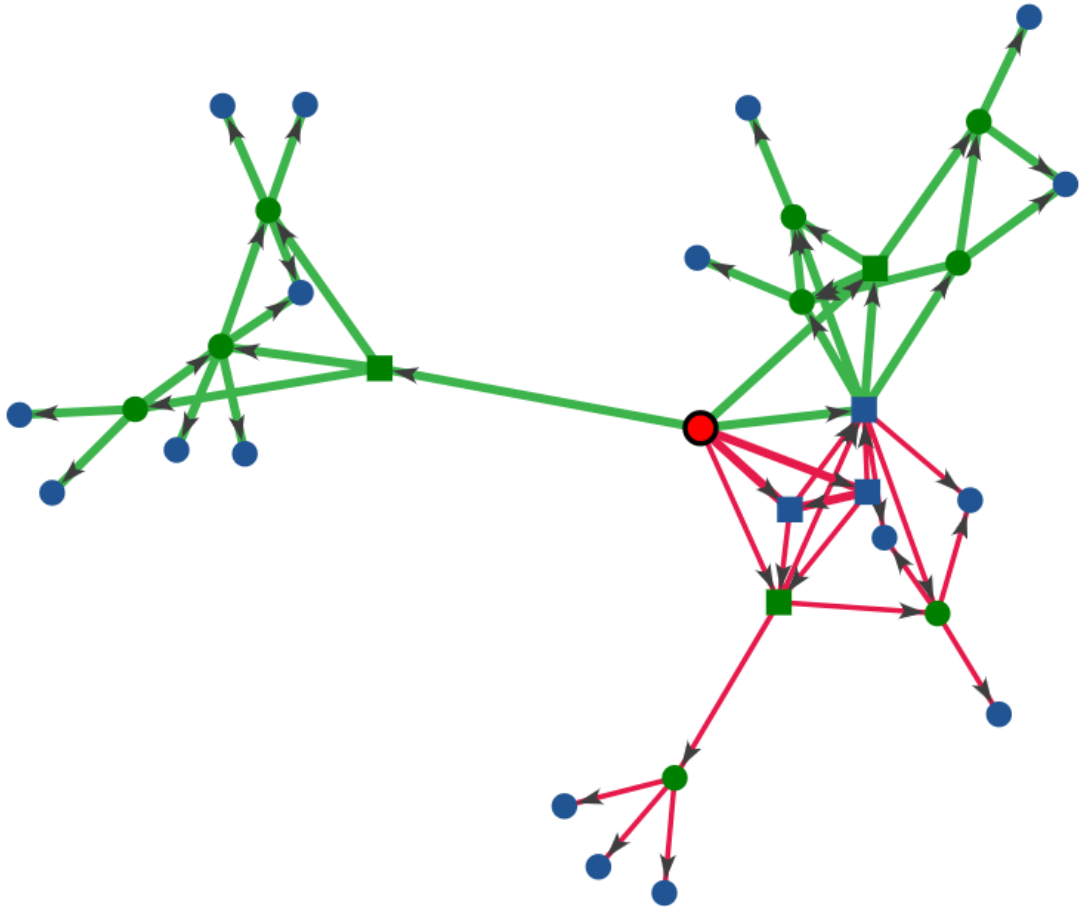
PhysChem Data X



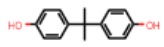
EPA United States Environmental Protection Agency

GenRA v3.1

Neighborhood Exploration Graph



Chemical Information



Name: Bisphenol A
Mol. Weight: 228.291
DTXCID: DTXCID30182
DTXSID: DTXSID7020182

Focus Chemical

Fingerprints

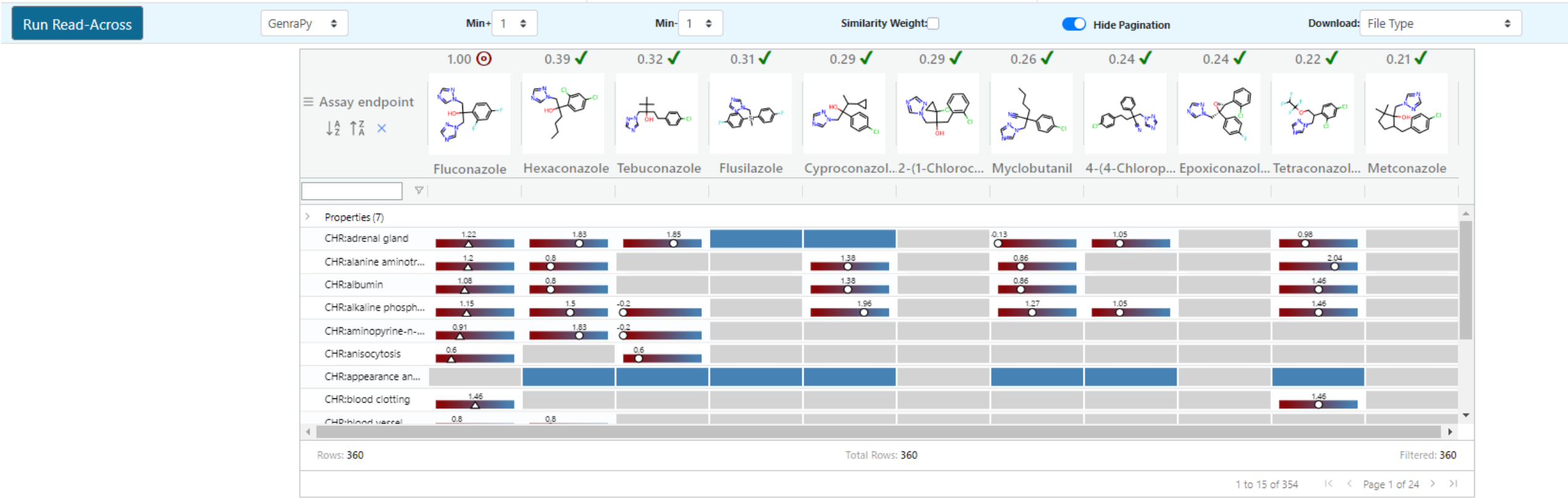
- Chem: Morgan Fgrprts
- Chem: Torsion Fgrprts
- Chem: ToxPrints
- Biology: ToxCast data
- Biology: ToxCast data, ATG
- Biology: ToxCast data, BSK
- Biology: ToxCast data, NVS
- Toxicity: ToxRef data
- Biology: HTPP_U2OS

Graph Type: all-paths

Filter By: ToxRef data

EPA United States Environmental Protection Agency

GenRA v3.1



GenRA next release tbd

- Speed enhancements
- Change download file to allow easier sorting and ranking based on AUC and p-value
- Incorporate new sorting and filtering in Panel 4
- Analogue Identification Methodology (AIM) fingerprints (Adams et al, under review)
- Download top 100 chemicals and their fingerprint representations
- Sync underlying data sources to accommodate recent updates to ToxCast data

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

EPA United States Environmental Protection Agency

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 established a baseline in performance. The approach relied 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 v3.0 released is a standalone web app linked to the Dashboard. A python package (genra-py) was released (March 2021) to facilitate batch processing using user specific datasets.
- Latest release is GenRA v3.1

EPA United States Environmental Protection Agency

Questions ?