

Generalized Read-Across (GenRA)

Overview

Read-across is a well-established data gap filling technique that is used within analogue and category approaches for regulatory purposes. Read-across represents the application of data from a source chemical(s) for a particular property or effect to predict the same property or effect for the target chemical (the chemical of interest) (OECD, 2014).

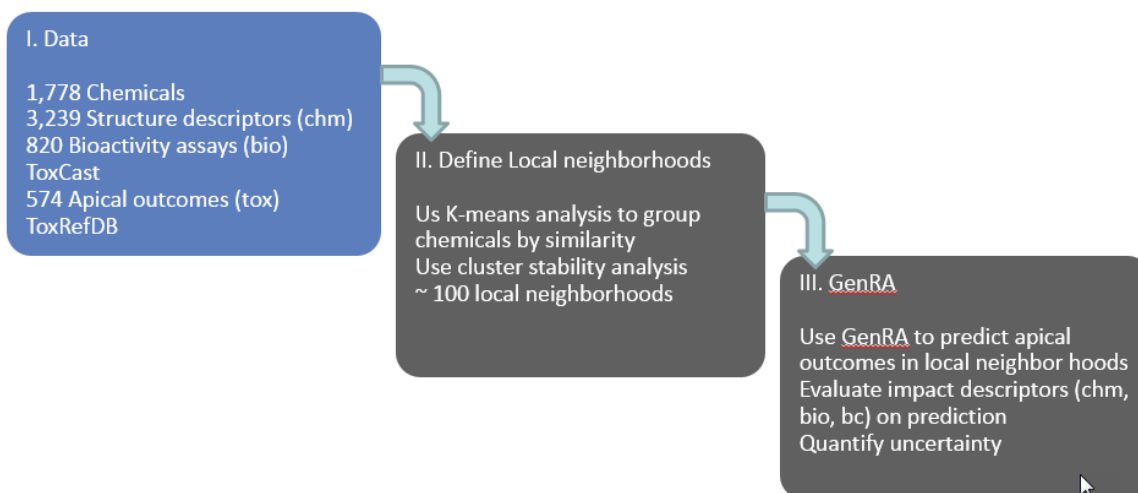
Here we present an implementation of an algorithmic automated approach to make reproducible read-across predictions of toxicity outcomes from *in vivo* studies called Generalized Read-across (GenRA) (Shah et al., 2016). The read-across prediction is a similarity weighted activity of nearest neighbors (source chemicals) based on chemistry and/or bioactivity descriptors. The approach is a generalization of the Chemical Biological Read-Across (CBRA) approach published by Low et al (2013). GenRA has been described in more detail in the literature (see Shah et al., 2016; Helman et al., 2018). Here we outline the principles of the approach and its implementation in the EPA CompTox Chemicals Dashboard.

Background

The GenRA approach was developed using available chemistry descriptor information, bioactivity High Throughput Screening (HTS) data from the ToxCast program and *in vivo* toxicity data from ToxRefDB v1.0 (see Figure 1). These experimental data are publicly available at <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcastm-data>

Chemical descriptor information comprising Morgan fingerprints (Rogers and Hahn, 2010) and topological torsion descriptors (Nilakantan et al., 1987) were generated for a set of 1778 chemicals taken from ToxCast Phases I-III. Bioactivity descriptors (denoted bio) comprised hit calls (active (1), inactive (0)) from 820 ToxCast HTS assays. These were used either singly (chm or bio to denote either chemical or bioactivity descriptors) or together (hybrid descriptor sets of both chemical and bioactivity, denoted as bc) to predict different toxicity outcomes from over 10 different study types from ToxRefDB v1.0.

Figure 1: Development of the GenRA approach



The study types included chronic, subacute, multigenerational, developmental guideline or guideline type studies (see Figure 2). Figure 2 provides a representation of the distribution of positive and negative outcomes across the different toxicity effects.

Figure 2: Study types represented in ToxRefDB v1.0



Chemicals were first clustered into predefined groups or neighborhoods. The GenRA algorithm was then used to make predictions within these neighborhoods (categories) for the different toxicity outcomes. Receiver operating characteristic (ROC) analysis was conducted for k-nearest neighbors (where the value of k ranged from 1 to the maximum number of chemicals in the neighborhood), and with a similarity threshold, s (where the value of s ranged from the minimum to maximum values of s across all unique pairwise comparisons in the neighborhood). The area under the curve (AUC) was then taken as a measure of performance for a given k and s value.

Category/Analogue workflow

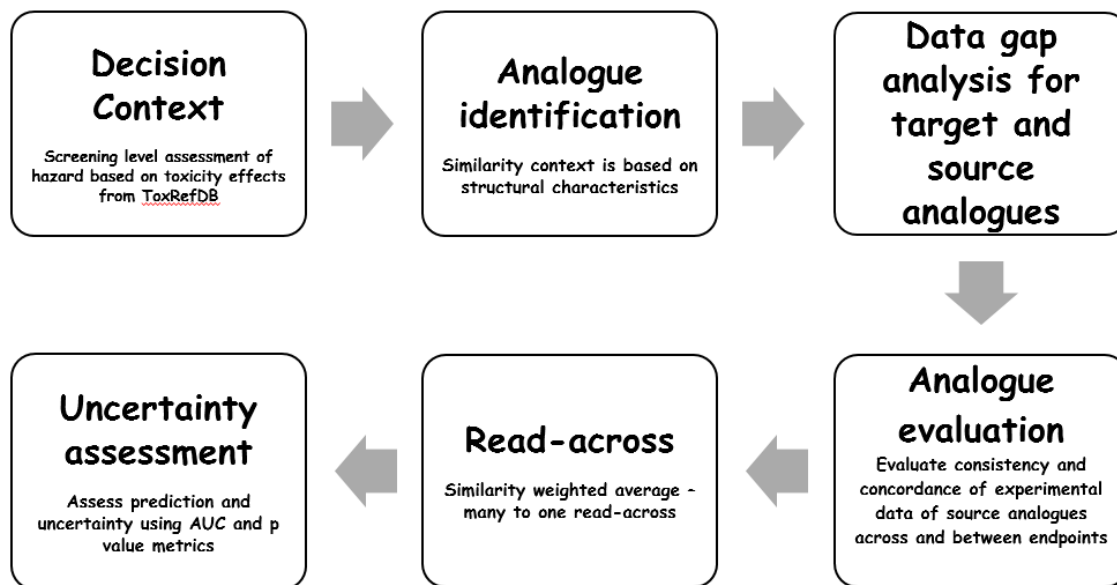
There are a number of steps in the development of a category or analogue approach. The seven key steps in the workflow are as follows:

1. Decision context
2. Data gap analysis
3. Overarching similarity rationale
4. Analogue identification
5. Analogue evaluation
6. Data gap filling
7. Uncertainty assessment

For more information, describing each of these steps in turn, see Patlewicz et al (2017; 2018).

In the GenRA implementation, the steps have been addressed are shown in Figure 3 (Helman et al., 2018). These will be illustrated using an example case study and walking through the various steps in the Dashboard implementation. The use case implemented addresses a qualitative prediction of a target chemical.

Figure 3: Category/Analogue workflow and GenRA

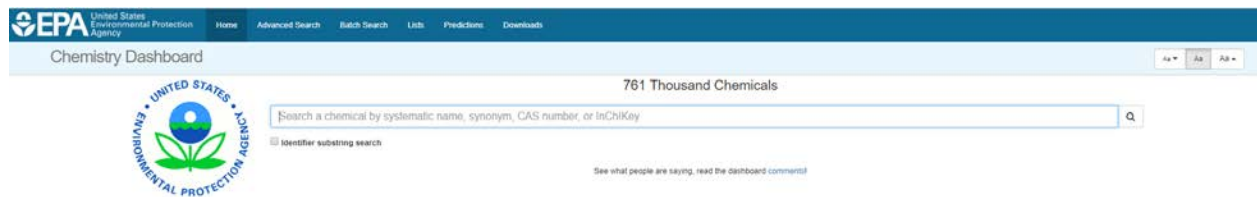


The starting point for GenRA relies on identifying the chemical of interest (target chemical) by performing a basic search within the EPA CompTox Chemicals Dashboard.

Basic Search

There are a variety of search capabilities presently available on the Dashboard including by chemical name or identifier. The text search box (Figure 4) allows a user to search using a number of chemical "identifiers" including chemical name, common name, [CAS Number](#) or [InChIKey](#).

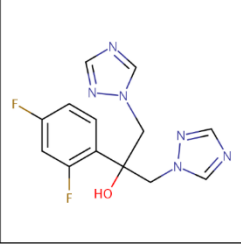
Figure 4: Text search box within the Dashboard



If a hit is identified in the database then the search will return a detailed results page with associated information for the chemical. For example, a search for "Fluconazole" will return a Chemical Results Page (Figure 5).

Figure 5: Chemical Results Page for Fluconazole

Fluconazole
86386-73-4 | DTXSID3020627
Searched by DSSTox Substance Id.



Wikipedia
Fluconazole is an antifungal medication used for a number of fungal infections. This includes candidiasis, blastomycosis, coccidioidomycosis, cryptococcosis, histoplasmosis, dermatophytosis, and pharyngeal candidiasis. 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

Intrinsic Properties

- Molecular Formula: $C_{12}H_{12}F_2N_4O$ [Mol File](#)
- 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

The page (Figure 5) shows an image of the 2D chemical structure and associated information.

An active tab (denoted by the black hyperlink and marked in red in Figure 5) named GenRA should be visible alongside the other data streams available on the Dashboard.

GenRA

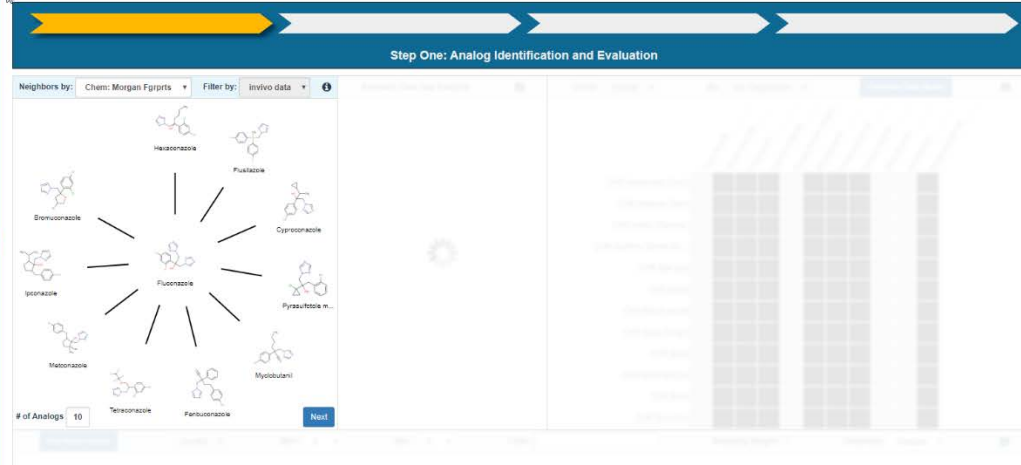
Once the GenRA tab is clicked, a grid like display is presented as shown in Figure 6. A workflow indicator above the display denotes the relevant step in the workflow.

Figure 6: Working interface of GenRA

Fluconazole
86386-73-4 | DTXSID3020627
Searched by DSSTox Substance Id.

Step One: Analog Identification and Evaluation

Neighbors by: Chem: Morgan Fprpts | Filter by: in vivo data



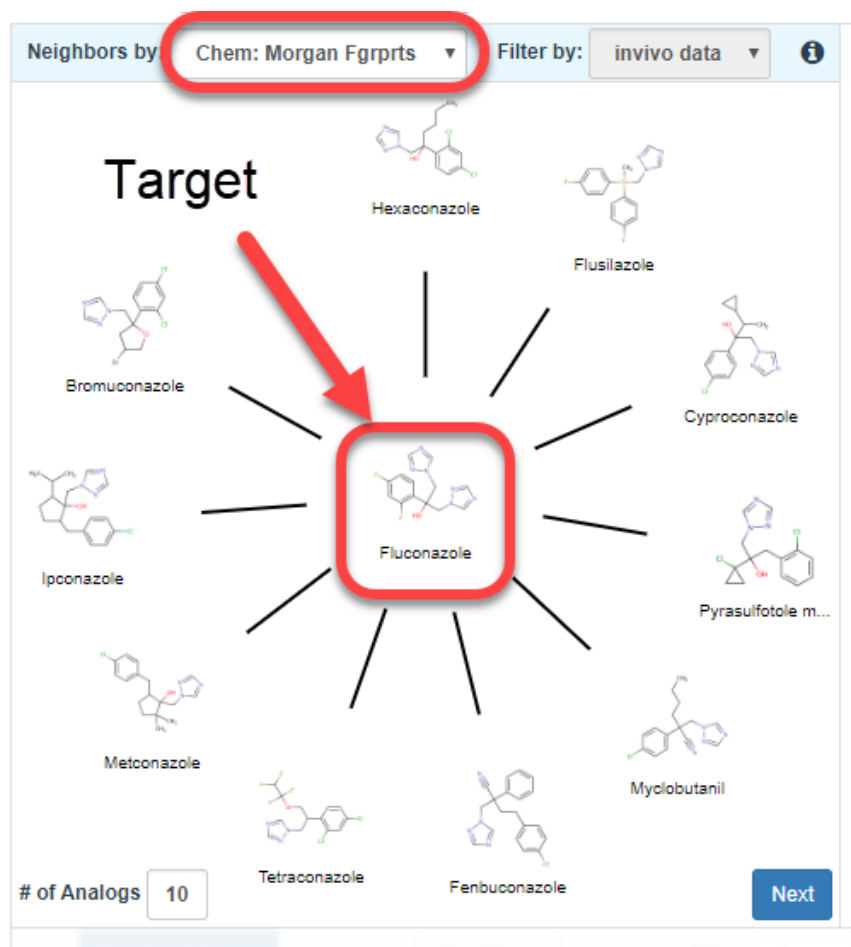
of Analogs: 10 [Next](#)

The category/analogue workflow as outlined in Figure 3 is then followed.

Analogue identification

This step involves searching for potential source analogues (nearest neighbours) based on some similarity context. This requires describing the target chemical using numeric representations of its structure and/or activity. There are different means of searching for source analogues – in Figure 7, Morgan chemistry fingerprints (Chem:Morgan Fgrprts) have been selected as the similarity context. The radial plot depicts 10 nearest neighbors (# of Analogs) filtered by availability of *in vivo* data from ToxRefDB v1.0 (invivo data).

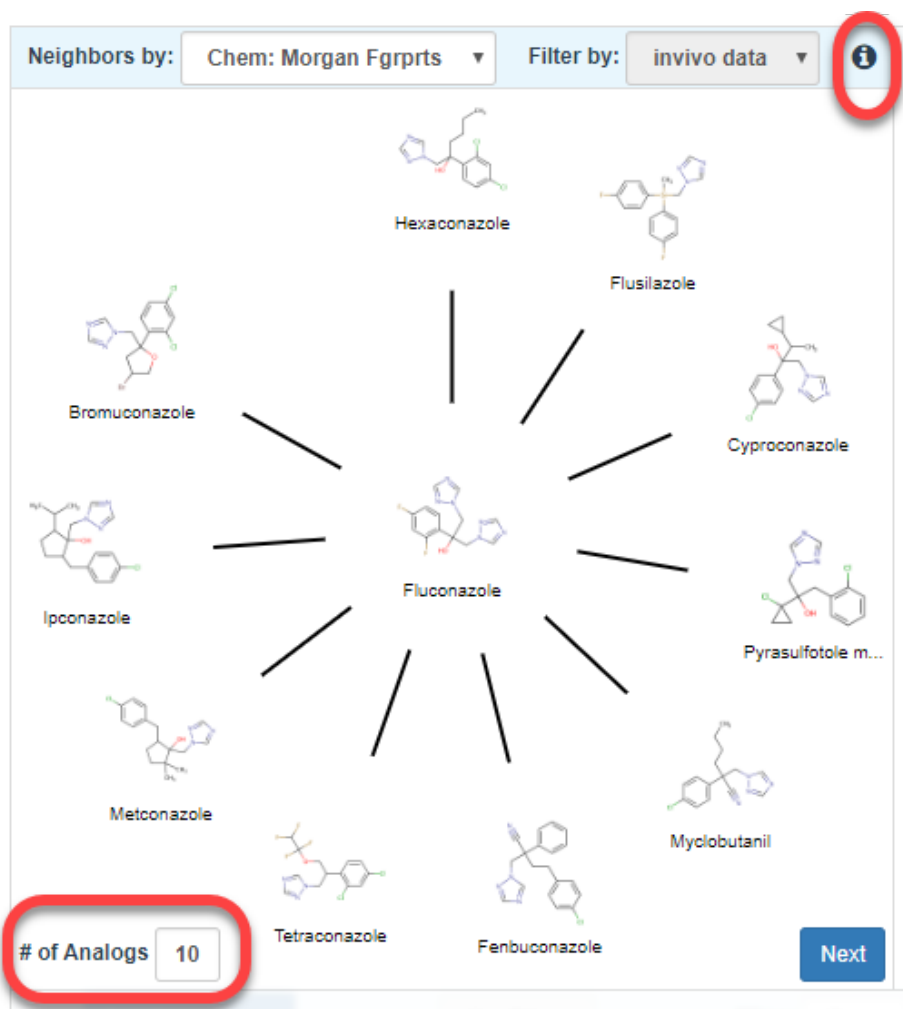
Figure 7: Analogue identification by Morgan fingerprints



These defaults can be changed to permit an update of the nearest neighbour (source analogue) radial plot to show k nearest neighbors according to a different similarity context whether that be, the k-nearest neighbors on the basis of a

different chemical fingerprint descriptor such as torsion fingerprints (Chem: Torsion Fgrprts), ToxPrints (Yang et al., 2015) or indeed bioactivity descriptors such as ToxCast hitcalls (Biology: ToxCast) or Tox21 hitcalls. The number of source analogues can be changed up to a maximum number of 15. A minimum number of source analogues is currently set to 5. As an end-user, it is typically more helpful to identify source analogues that are associated with relevant *in vivo* data to permit a read-across prediction to be made. In the current implementation, analogues are automatically filtered on the basis of ToxRefDB *in vivo* data. In future versions, the ability to filter on the basis of different data sources such as *in vitro* data from ToxCast will also be made available. The “i” icon provides help for what aspects are captured in the first step of the workflow (Figure 8). These help icons are replicated in all the other grids in the interface.

Figure 8: Changing the number of analogues and the help function

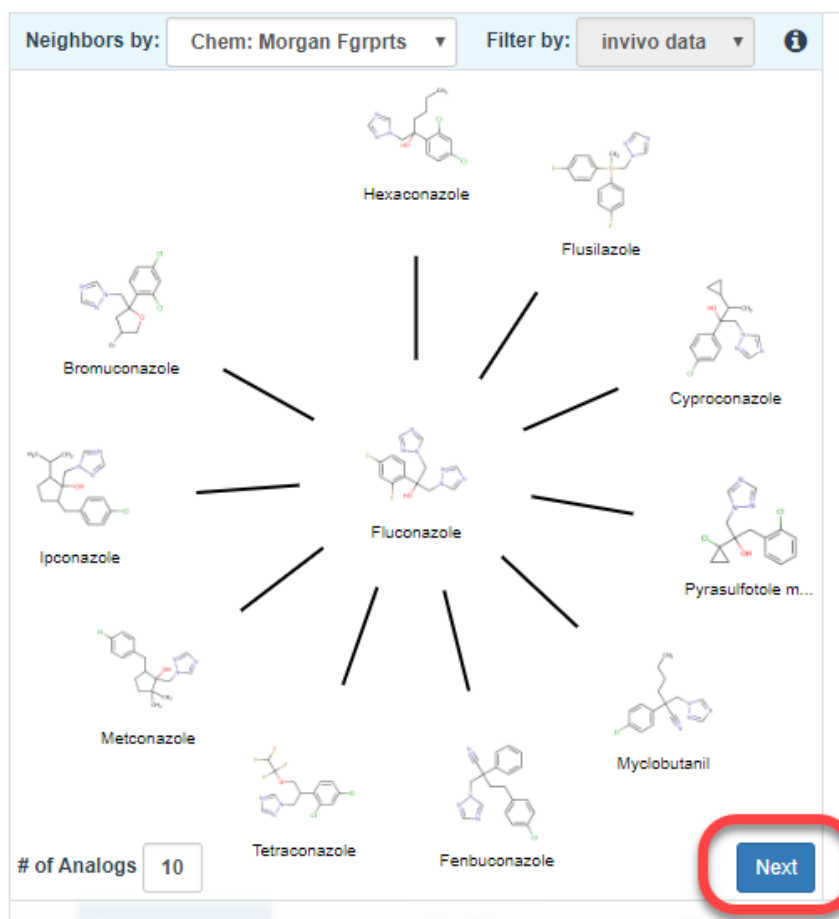


The radial plot depicts the source analogues in decreasing order of similarity using the Jaccard similarity metric. This similarity metric goes between 0 – 1 where 1 denotes the same and 0, dissimilar. No specific thresholds are set on how similar an analogue ought to be for it to be included in the analysis. Hovering over any of the analogues in the plot will highlight them in turn and depict the similarity index as a number. In Figure 7, using Morgan fingerprints as descriptors, the pairwise similarity between Hexaconazole and Fluconazole is 0.39 whereas the pairwise similarity for Fluconazole and Bromuconazole is 0.2. The subscript of c denotes that the Jaccard similarity is taking into account chemical features. The subscript would be b for biological descriptors.

If the user wishes to conduct a GenRA for a different source analogue represented in the radial plot, clicking on any of the analogues will open a new Chemical Results tab for that source analogue in the browser.

Once the user is satisfied with the number of analogues, clicking on the “Next” button as shown in Figure 9 proceeds to the next step of the workflow.

Figure 9: Proceeding to the Data Gap Analysis step in the workflow



Data gap Analysis

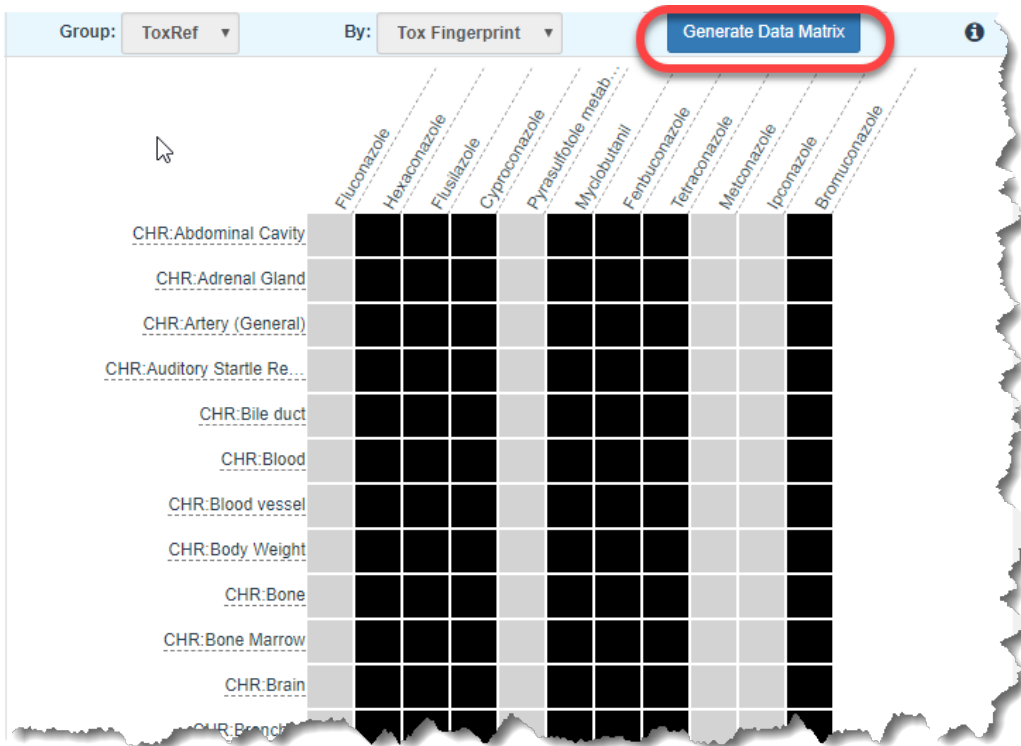
A summary overview of the available data quantity for the target chemical Fluconazole and its source analogues is provided in the second grid (Figure 10). The color density represents the data availability for the target – from light to dark. A light grey box represents no data whereas black represents the greatest number of records. The number of data records is reflected in the box itself. The data availability is segmented by data type – where bio_tx21 represents bioactivity data from Tox21, bio_txct represents bioactivity data from ToxCast, chm_ct represents chemistry descriptor information and tox_txrf represents *in vivo* toxicity effect information from ToxRefDB v1.0. In the case of target Fluconazole, there are 'some' bioactivity and chemistry data (based on the lighter colors in the boxes) from which toxicity predictions could be made but there are no *in vivo* toxicity data available in ToxRefDB v1.0.

Figure 10: Summary Data Gap Analysis view

	bio_tx21	bio_txct	chm_ct	tox_txrf
Fluconazole	3	714	15	0
Hexaconazole	43	819	18	345
Flusilazole	28	819	9	345
Cyproconazole	14	819	16	408
Pyrasulfotole metabolite ...	0	0	18	234
Myclobutanil	15	818	15	345
Fenbuconazole	34	819	17	345
Tetraconazole	35	819	20	345
Metconazole	35	215	15	82
Ipcconazole	46	232	16	180
Bromuconazole	24	277	13	345

The third grid (Figure 11) represents the available data captured on the basis of the toxicity effects within the ToxRefDB v1.0 studies. Here a box marked in black indicates the availability of information vs lack of any information. Once the user has browsed the matrix to identify what types of data gaps exist and the extent to which these might be filled by the source analogues identified based on their existing data, the button “generate data matrix” is clicked to derive a data matrix view that summarizes the same information but on the basis of activity score (presence or absence of toxicity effects).

Figure 11: Summary Data Gap analysis view on the basis of toxicity effects



Analogue evaluation

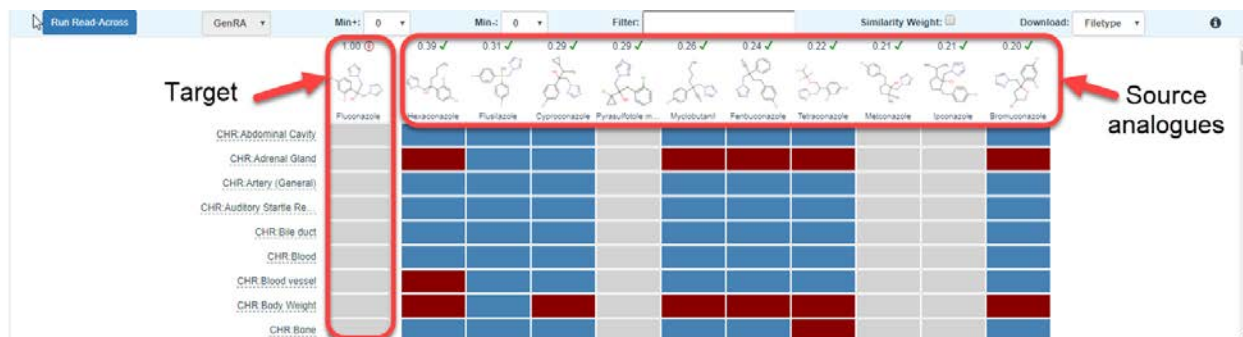
Analogue evaluation entails evaluating the suitability of the source analogues identified. Part of this evaluation already involves looking at the data availability of those source analogues (a data gap analysis across the source analogues and the target). If little data are available for the source analogues or they fail to address the data gaps of interest for the target chemical then this might lead the user to change the number of neighbors or select a different similarity context. Grid windows 2 and 3 in Figures 10 and 11 provide a context of available data for the source analogues both in terms of the quantity of data and its type as well as across study type on the basis of the toxicity effects.

In Figure 11, the Tox fingerprint reflects the toxicity effects within each study type. Since there are some 129 different toxicity effects represented in ToxRefDB v1.0, the user should browse these effects by using the scroll bar.

For a more detailed evaluation of the source analogues in order to evaluate their concordance and consistency within and across the study types, the Generate Data Matrix button needs to be clicked as shown in Figure 12. This produces

a data matrix view. Across the top are the target chemical, the source analogues ordered by similarity. Each row represents a toxicity effect.

Figure 12: Data matrix view

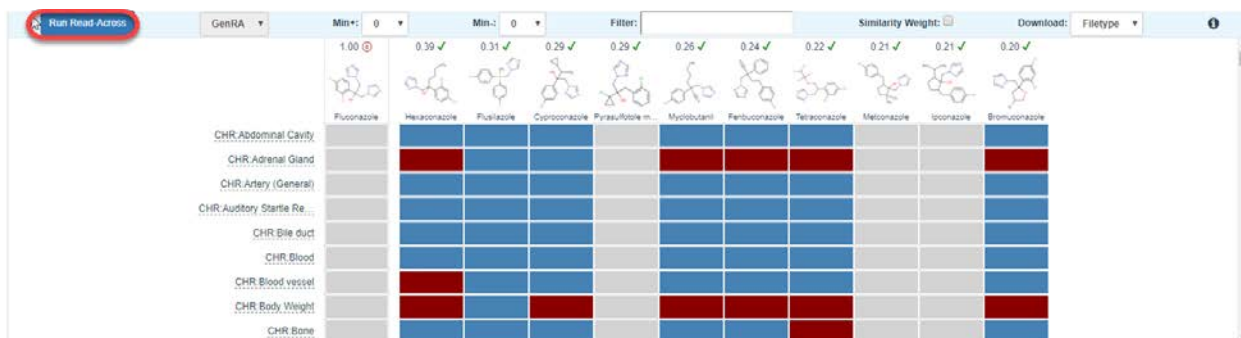


The toxicity effects are populated by red and blue boxes across the analogues representing the presence or absence of toxicity effects in the *in vivo* studies from ToxRefDB v1.0. Blue indicates an absence of effects whereas red represents presence of toxicity effects. Grey boxes indicate an absence of information. Hovering over the red boxes will show the dose at which a toxicity effect was observed. This data matrix view for the source analogues enables a quick perspective to evaluate the suitability of the analogues and the trends they exhibit in terms of their toxicity effects. This allows data gaps to be readily identified.

Data gap filling

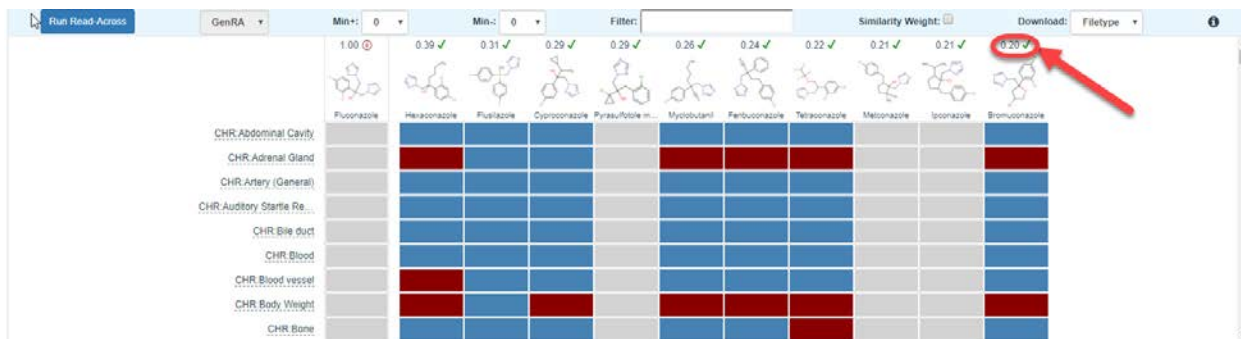
This step is where the GenRA prediction is generated. A prediction can be generated by clicking on “Run Read-across” as indicated in Figure 13.

Figure 13: Data gap filling by GenRA



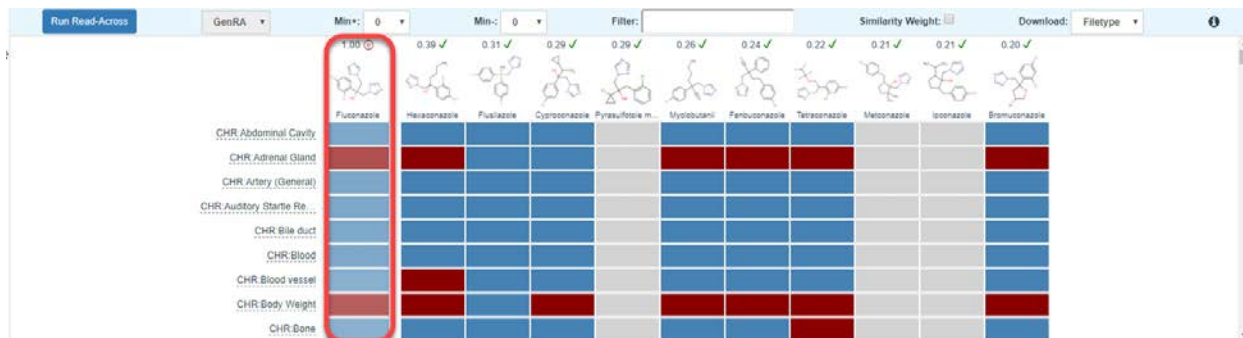
Predictions will by default be run on the screen visible. The toxicity effects being predicted can also be filtered using the “Filter” box. The predictions can also be modified by specifying thresholds for the number of actives and inactives across the analogues using the Min+ and Min- filters. The numbers above any source analogue reflects the similarity index. Clicking on the Similarity weight checkbox modifies the size of the box to reflect the pairwise similarity metric. After evaluating the source analogues, an end user can deselect an analogue if it is lacking in data or if upon expert review it is deemed to be an outlier in the overall trend of toxicity effects across the source analogues (Figure 14).

Figure 14: De-selecting analogues within the data matrix



Clicking on “Run Read-across” will update the target information to show the predictions made (Figure 15). The opacity of the predictions reflects the confidence in the prediction made with. A faint colored prediction will denote lower confidence in the prediction.

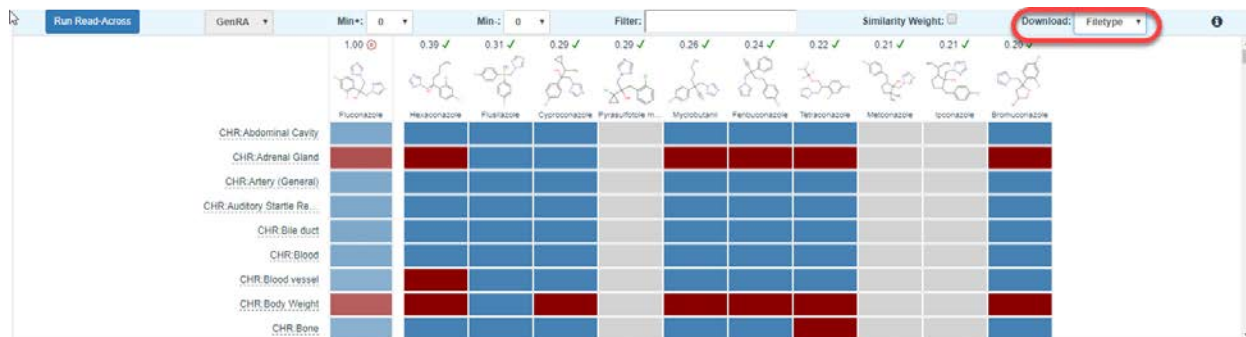
Figure 15: Data gap filling predictions by GenRA



Uncertainty assessment

Predictions made can be exported by clicking on the Download button shown in Figure 16.

Figure 16: Exporting results from GenRA



The AUC performance measure is noted as part of the prediction provided. This can be interrogated when the predictions generated are download as a TSV file or Excel file. A snapshot of what the download file resembles is depicted in Figure 17.

Figure 17: Sample prediction output produced as an export

cls	target	analog	analog	analog	analog	analog	analog	analog	analog	analog	analog	analog
label	Fluconazole	Hexaconazole	Flusilazole	Cyproconazole	Pyrasulfotole (SXX 0665)	Myclobutanil	Fenbuconazole	Tetraconazole	Metconazole	Ipconazole	Bromucrole	
dsstox_cid	DTXCID10627	DTXCID2014 653	DTXCID70 4235	DTXCID8012 601	DTXCID1024 338	DTXCID3043 15	DTXCID6012 548	DTXCID6014 956	DTXCID2014 497	DTXCID5014 674	DTXCID 531	
casrn	86386-73-4	79983-71-4	85509-19-9	94361-06-5	120983-64-4	88671-89-0	114369-43-6	112281-77-3	125116-23-6	125225-28-7	116255	
jaccard		0.38888889	0.3125	0.28947368	0.28571429	0.25641026	0.23809524	0.22352941	0.21176471	0.20689655	0.1976	
CHR:Abdominal Cavity	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Adrenal Gland	GenRA Pos Act=1 (0.684) AUC=0 p=0.98	4.700 mg/kg/day	no_effect	no_effect	no_data	70.200 mg/kg/day	30.000 mg/kg/day	27.700 mg/kg/day	no_data	no_data	25.000 mg/kg/day	
CHR:Artery (General)	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Auditory Startle Reflex Habituation	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Bile duct	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Blood	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Blood vessel	GenRA Neg Act=0 (0.204) AUC=0 p=0.895	50.000 mg/kg/day	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Body Weight	GenRA Pos Act=1 (0.836) AUC=0 p=0.89	6.100 mg/kg/day	no_effect	3.200 mg/kg/day	no_data	6.250 mg/kg/day	3.750 mg/kg/day	12.970 mg/kg/day	no_data	no_data	55.600 mg/kg/day	
CHR:Bone	GenRA Neg Act=0 (0.117) AUC=0 p=0.89	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	27.700 mg/kg/day	no_data	no_data	no_eff	
CHR:Bone Marrow	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Brain	GenRA Neg Act=0 (0.355) AUC=0 p=0.98	no_effect	no_effect	no_effect	no_data	393.500 mg/kg/day	no_effect	59.000 mg/kg/day	no_data	no_data	134.5 mg/kg/day	
CHR:Bronchus	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	
CHR:Cervix	GenRA Neg Act=0 (0) AUC=0 p=1	no_effect	no_effect	no_effect	no_data	no_effect	no_effect	no_effect	no_data	no_data	no_eff	

Here a user can see the predictions made, the AUC and the p-value that is associated with the prediction. In this case the p-values are high indicating that the confidence in the predictions are not considered significant. The actual experimental data for the source analogues is also reflected.

References

- Helman G, Shah I, Patlewicz G. 2018. Extending the Generalised Read-Across approach (GenRA): A systematic analysis of the impact of physicochemical property information on read-across performance. *Comp Toxicol. In press*
- Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A. 2013. Integrative chemical-biological read-across approach for chemical hazard classification. *Chem. Res. Toxicol.* 26(8): 1199-1208.
- Nilakantan R, Bauman N, Dixon JS, Venkataraghavan R. 1987. Topological torsion - a new molecular descriptor for SAR applications - comparison with other descriptors. *J. Chem. Infor. Comput. Sci.* 27(2): 82-85.
- OECD. 2014. Guidance on grouping of chemicals. OECD Series on Testing and Assessment No. 194. Organisation for Economic Co-operation and Development, Paris, France.
- Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G. 2016. Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. *Regul. Toxicol. Pharmacol.* 79: 12-24. doi: 10.1016/j.yrtph.2016.05.008.
- Patlewicz G, Helman G, Pradeep P, Shah I. 2017. Navigating through the minefield of read-across tools: A review of in silico tools for grouping. *Comp. Toxicol.* 3: 1-18
- Patlewicz G, Cronin MTD, Helman G, Lambert J, Lizarraga LE, Shah I. 2018. Navigating through the minefield of read-across frameworks: A commentary perspective. *Comp. Toxicol.* 6: 39-54
- Rogers D, Hahn M. 2010. Extended-connectivity fingerprints. *J. Chem. Infor. Model.* 50: 742-754.

Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G. 2016. Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. *Regul Toxicol Pharmacol.* 79: 12-24.

Yang C, Tarkhov A, Marusczyk J, Bienfait B, Gasteiger J, Kleinoeder T, Magdziarz T, Sacher O, Schwab CH, Schwoebel J, Terfloth L, Arvidson K, Richard A, Worth A, Rathman J. 2015. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* 55(3): 510-528.