Generalised 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).
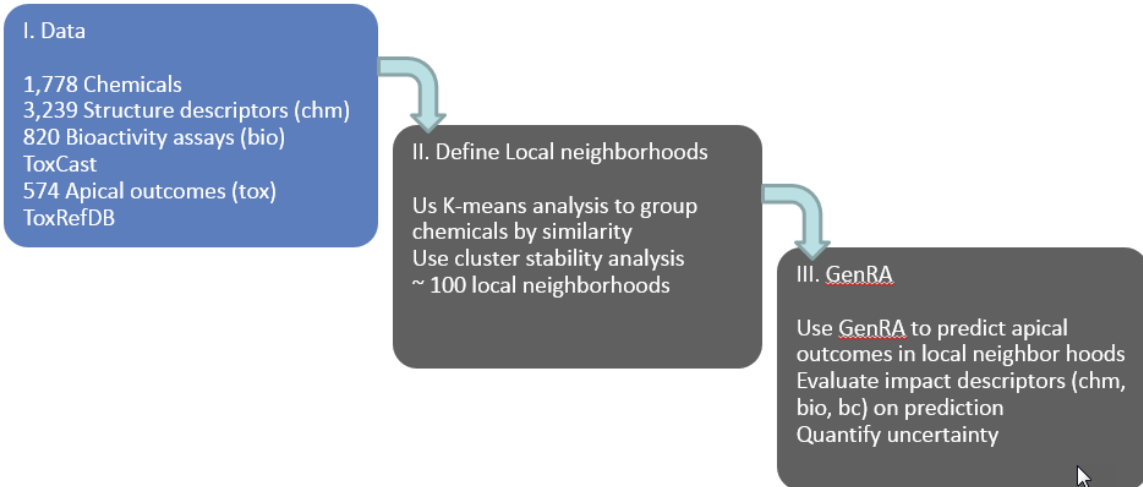
Here we present an implementation of an algorithmic automated approach to make reproducible read-across predictions of toxicity outcomes from *in vivo* studies called Generalised Read-across (GenRA) (Shah et al., 2016). The read-across prediction is a similarity weighted activity of nearest neighbours (source chemicals) based on chemistry and/or bioactivity descriptors. The approach is a generalisation 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; 2019; Patlewicz et al., 2018). Here we outline the principles of the approach and its implementation as a standalone web based application that is linked to the EPA CompTox Chemicals Dashboard.

Background

The GenRA approach was first developed using available chemistry descriptor information, bioactivity High Throughput Screening (HTS) data from the ToxCast program and *in vivo* toxicity data from ToxRefDB v.1.0 (see Figure 1). These experimental data are publicly available at https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-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

I. Data

1,778 Chemicals
3,239 Structure descriptors (chm)
820 Bioactivity assays (bio)
ToxCast
574 Apical outcomes (tox)
ToxRefDB

II. Define Local neighborhoods

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

III. GenRA

Use GenRA to predict apical outcomes in local neighbor hoods
Evaluate impact descriptors (chm, bio, bc) on prediction
Quantify uncertainty

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
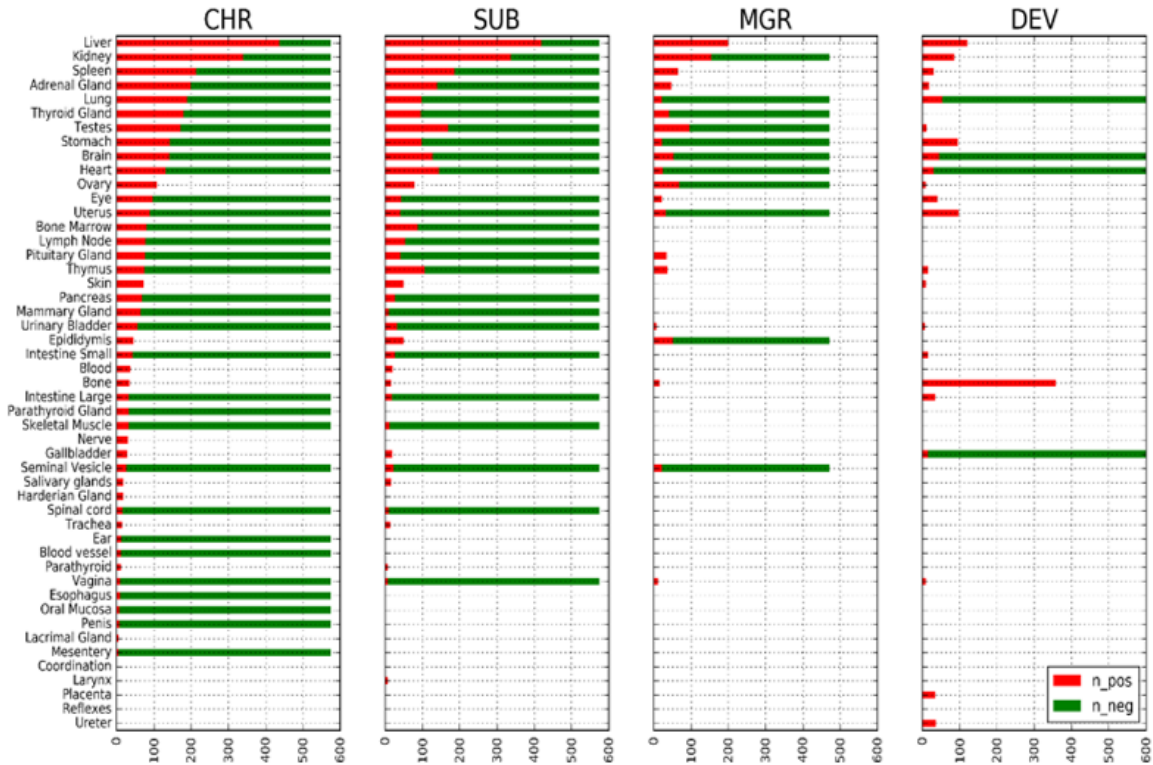
Chemicals were first clustered into pre-defined groups or neighbourhoods. The GenRA algorithm was then used to make predictions within these neighbourhoods (categories) for the different toxicity outcomes. Receiver operating characteristic (ROC) analysis was conducted for k-nearest neighbours (where the value of k ranged from 1 to the maximum number of chemicals in the neighbourhood), 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 neighbourhood). The area under the curve (AUC) was then taken as a measure of performance for a given k and s value.

Since this implementation, the most recent build of GenRA (Version 3 – released February 2022) is underpinned by *in vivo* toxicity data from ToxRefDB v2 (https://github.com/USEPA/CompTox-ToxRefDB) as well as the ToxCast/Tox21 data from invitrodb 3.4 (https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data).

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; 2019). These will be illustrated using an example case study and walking through the various steps in the webapp implementation. The use case implemented addresses a qualitative prediction of a target chemical.

Figure 3: Category/Analogue workflow and GenRA
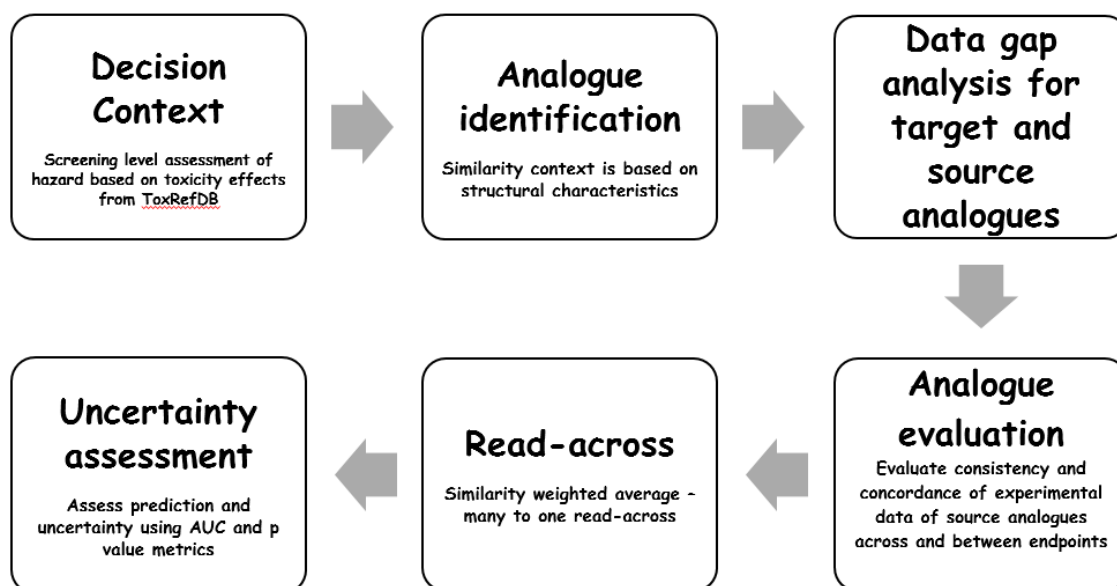
The starting point for GenRA relies on identifying a chemical of interest (target chemical). This can be performed in one of several ways depending on whether the route of entry is directly through the GenRA application or whether GenRA is accessed from within the EPA CompTox Chemicals Dashboard.

The most direct route of accessing the GenRA web application is through using the panel from the main CompTox portal (Figure 4). Navigate to comptox.epa.gov and click on the panel marked GenRA. This will redirect you to the main landing page of GenRA at https://comptox.epa.gov/genra/.

Figure 4: Landing page of the CompTox portal
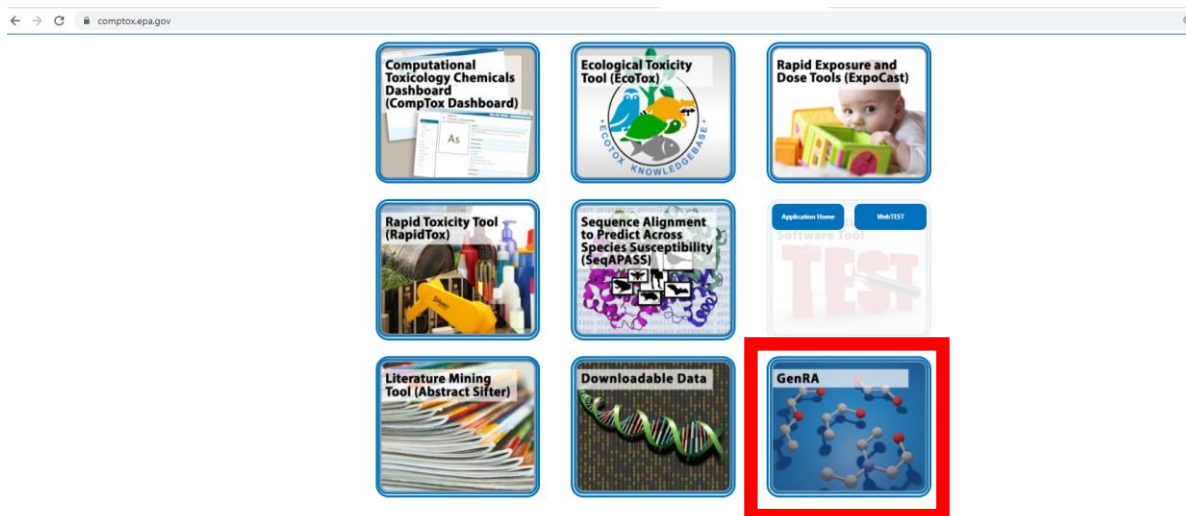
GenRA

Basic Search

There are a variety of search options to identify a target chemical using GenRA including chemical name, or other identifier. The text search box (Figure 5) allows a user to search using a number of these chemical "identifiers" including chemical name, common name, CAS Number, DSSTox identifier (DTXSID),InChIKey or SMILES (Simplified Molecular Input Line Entry System).

Figure 5: Text search box within the GenRA application
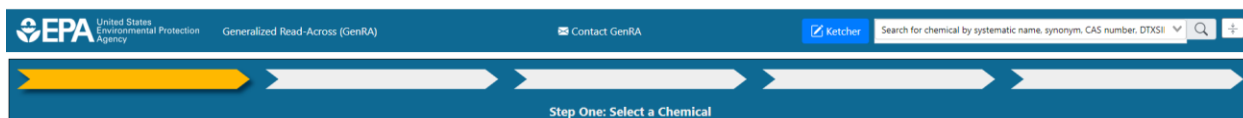
If a hit is identified in the database, then the search will return a grid like display where the workflow indicator above the display denotes the relevant step in the workflow. By default, a radial plot forms the first step where the target substance (for illustrative purposes Fluconazole is used) is shown in the centre and the 10 most similar analogues with associated ToxRefDB v2 data are shown clockwise in order of decreasing similarity (Figure 6)
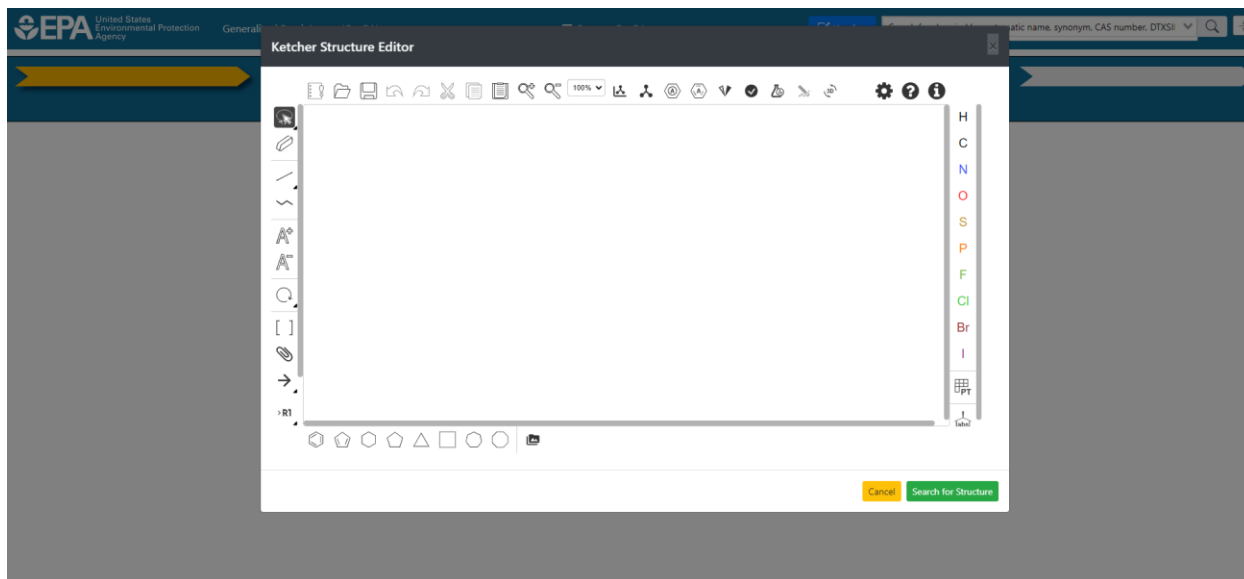
Figure 6: Radial plot returned for Fluconazole
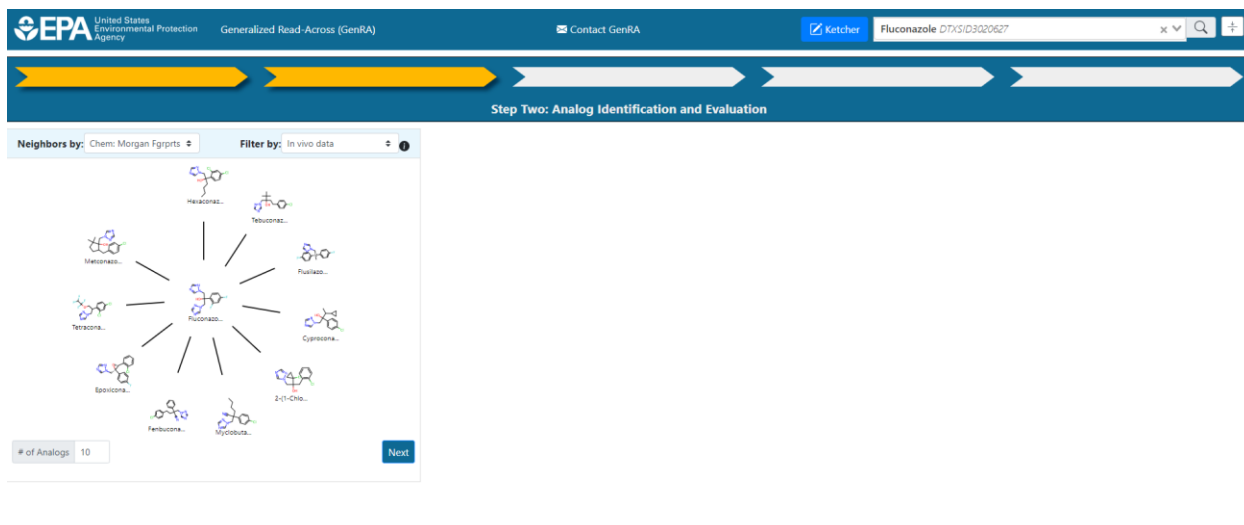
Advanced Search

An alternative means for searching for a target substance can be undertaken using the chemical drawing palette, Ketcher which is located left to the search text box (Figure 7).

Figure 7: Ketcher chemical drawing palette view when launched

Using the Ketcher drawing package, chemical structures can be introduced in a variety of ways. Chemical structures can be drawn freehand, SMILES accessed from elsewhere can be pasted into the palette window (click on the open

file icon and paste SMILES) or by importing a MOL file. Figure 8 shows how the SMILES for Fluconazole can be introduced into the palette.

Figure 8: Pasting SMILES into the Ketcher palette
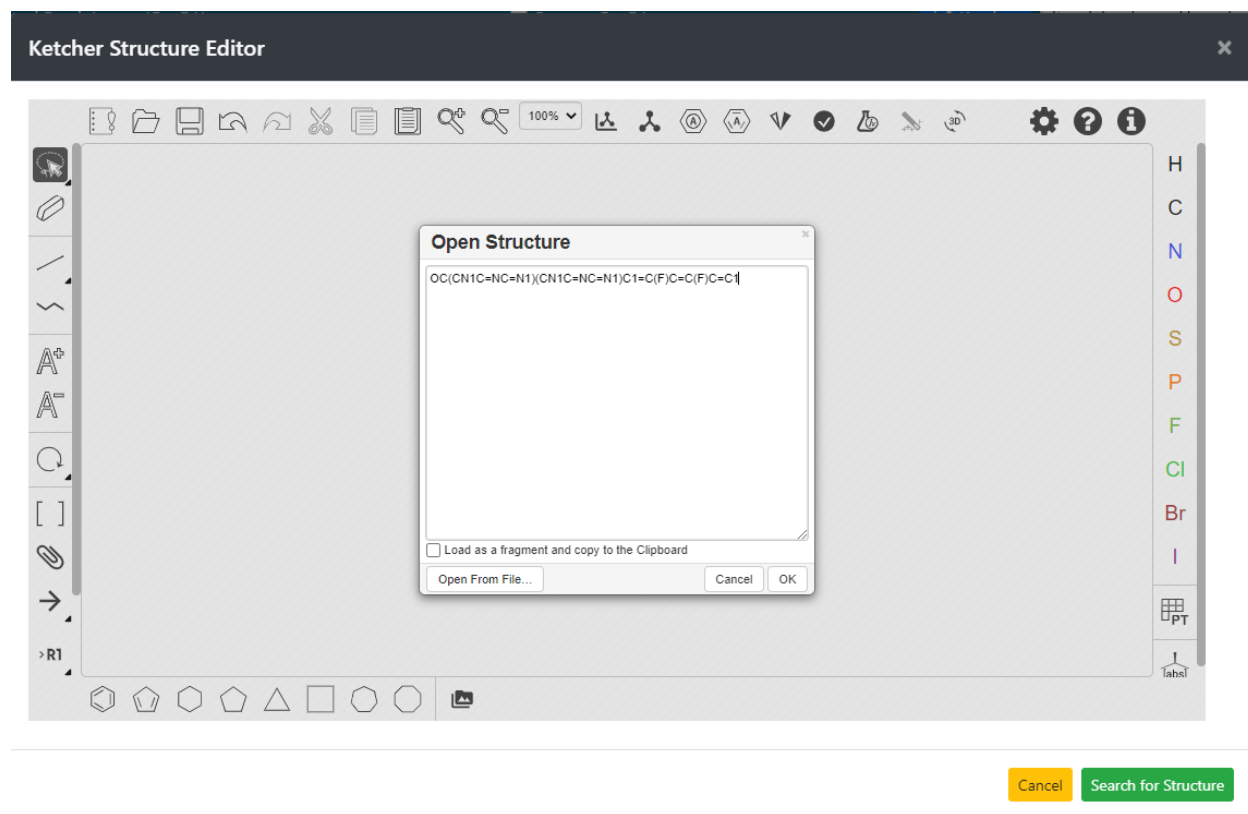
Once the structure representation is introduced, clicking on OK will show the structure in the palette itself. Clicking on 'Search for Structure' will then initiate the search to return the grid view in the same manner as when using the search text box. Note if a structure is used as the starting target – it is denoted as 'Unnamed chemical' in the resulting radial plot.

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 9, morgan chemistry fingerprints (Chem:Morgan Fgrprts) have been selected as the similarity context. The radial plot depicts the 10 nearest neighbours (# of Analogs) filtered by availability of *in vivo* data from ToxRefDB v2.0 (invivo data). The 10 nearest neighbours are shown in order of decreasing similarity. By default, Morgan fingerprints are used to return the 10 analogues and prefiltered based on *in vivo* data.

Figure 9: Analogue identification by Morgan fingerprints
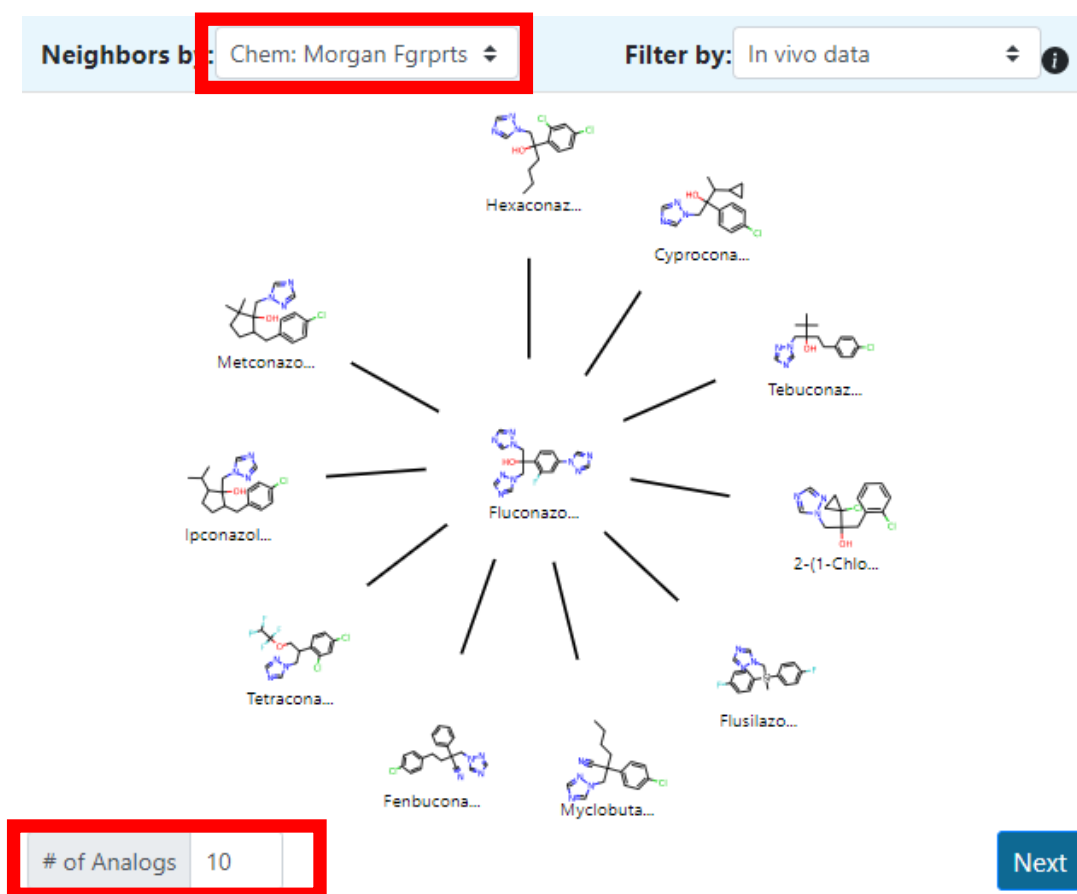
These defaults can be changed to permit an update of the nearest neighbour (source analogue) radial plot to show k nearest neighbours according to a different similarity context whether that be, the k-nearest neighbours on the basis of a different chemical fingerprint descriptor such as torsion fingerprints (Chem: Torsion Fgrprts), ToxPrints (Yang et al., 2015) or bioactivity descriptors such as ToxCast hitcalls (Biology: ToxCast). 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 1. 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. However, the end-user can update the view to show the most similar analogues regardless of available data. In this case, the filter option should be changed to 'No filter' to update the radial plot. Note this can be a slow process to update the resulting grid view (Figure 10).

Figure 10: Radial plot with no filter
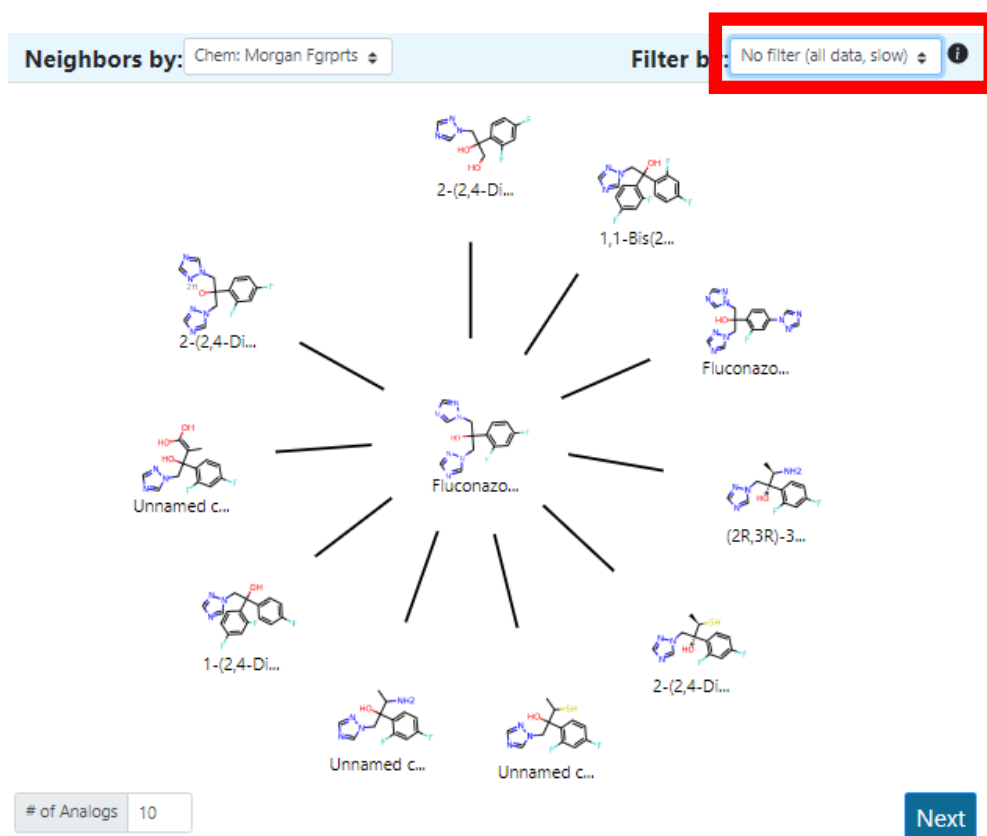
In addition to the standard fingerprint representations as described earlier, the end-user can also create custom fingerprint representations to search for analogues. Here the custom fingerprint option is selected (Figure 11) prompting the user to select up to 3 different fingerprints with associated weightings.

One or more of the chemical fingerprints and/or the ToxCast fingerprints can be selected adjusting the weights to make up a total of 100% e.g. 25% ToxPrints, 25% Morgan and 50% ToxCast fingerprints is shown in Figure 12. This will update the radial view to show the resulting 10 most similar analogues on the basis of these custom fingerprint representations (Figure 13).
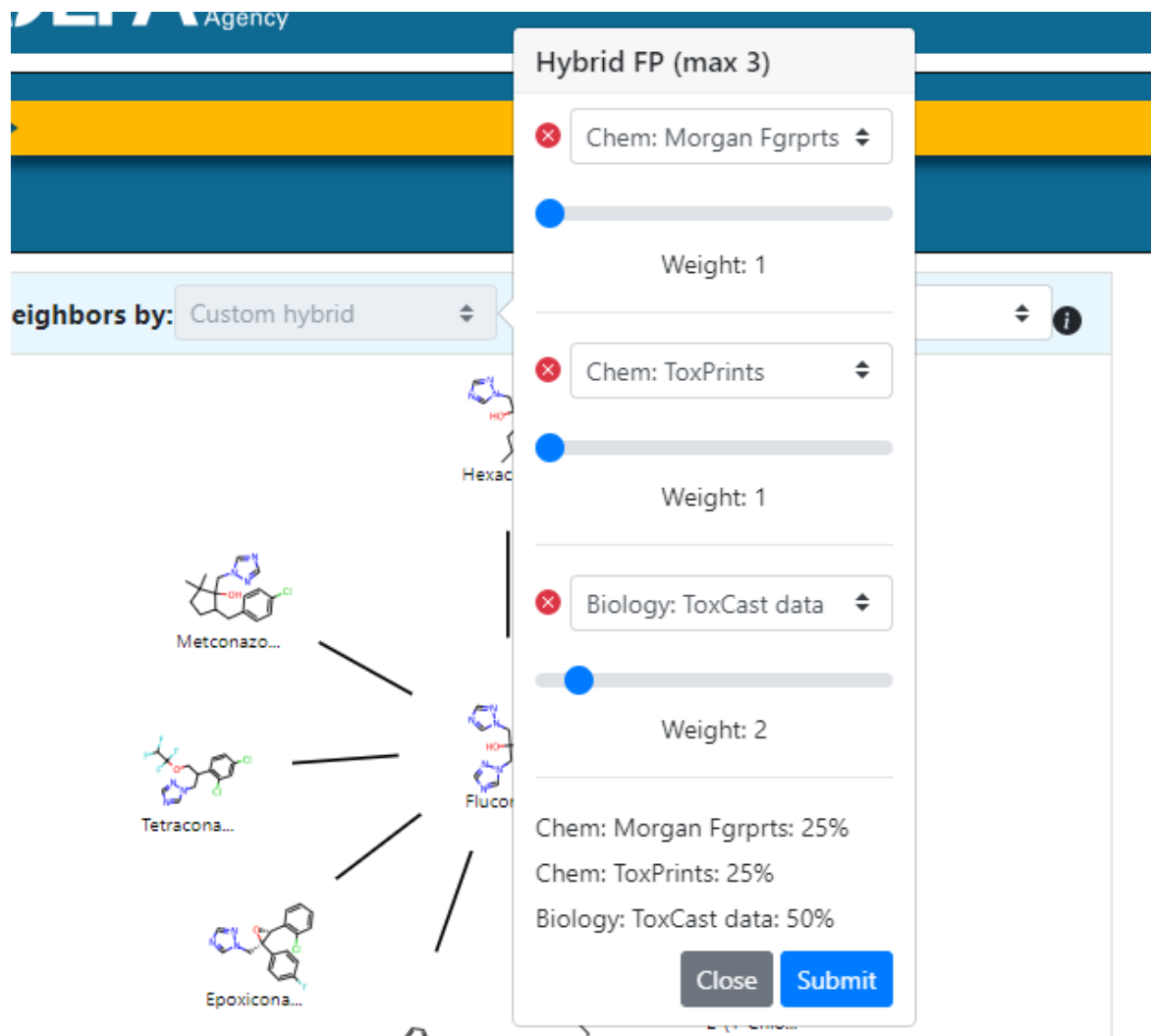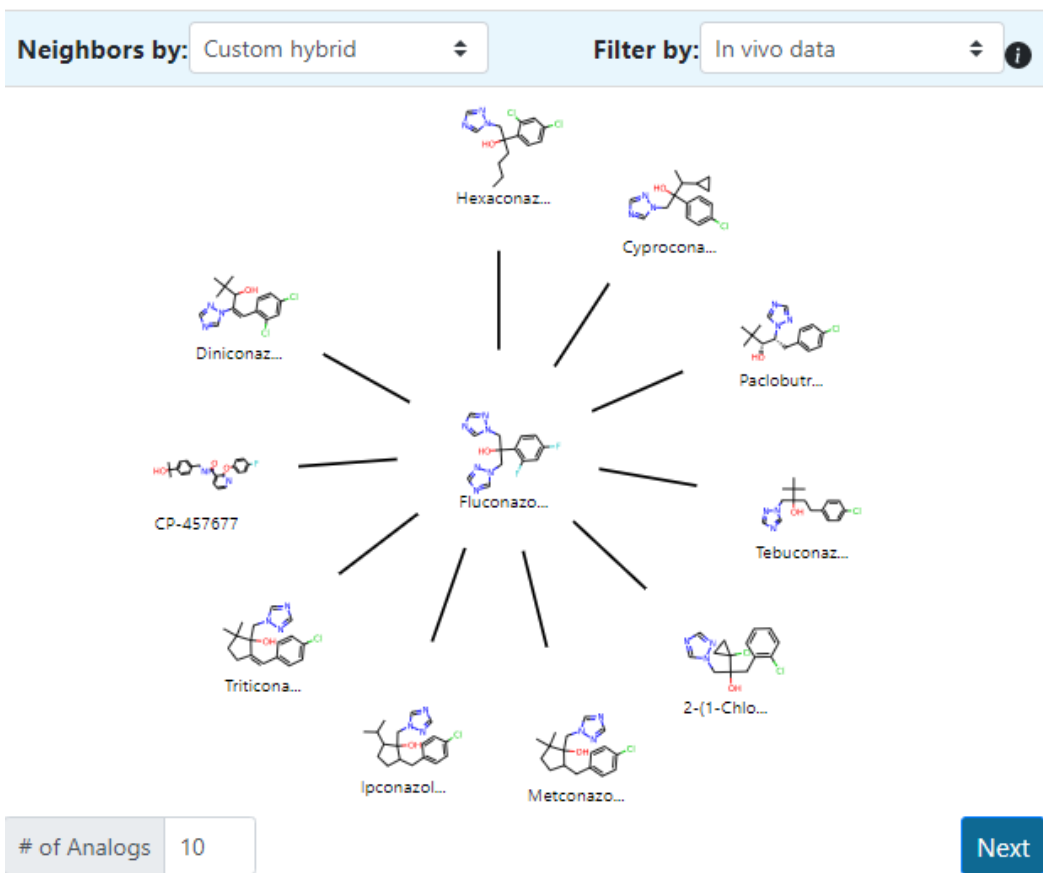
Figure 12: Custom fingerprint options

Figure 13: Updated radial plot using a custom fingerprint

The "i" icon provides help for what aspects are captured in the first step of the workflow (Figure 14). These help icons are replicated in all the other grids in the interface.

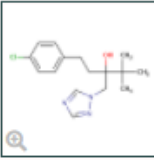Figure 14: Changing the number of analogues and the help function

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 Flusilazole is 0.31. The subscript of c denotes that the Jaccard similarity is taking into account chemical features. The subscript would be b for biological descriptors and x for custom descriptors.

Clicking on a different source analogue represented in the radial plot, will open a new Chemical Results tab in the CompTox Chemicals Dashboard for that source analogue. From there, the end-user can navigate to GenRA using the left hand panel and clicking on the relevant hyperlink (Figure 15).

Figure 15: Chemical landing page for the first source analogue of Fluconazole
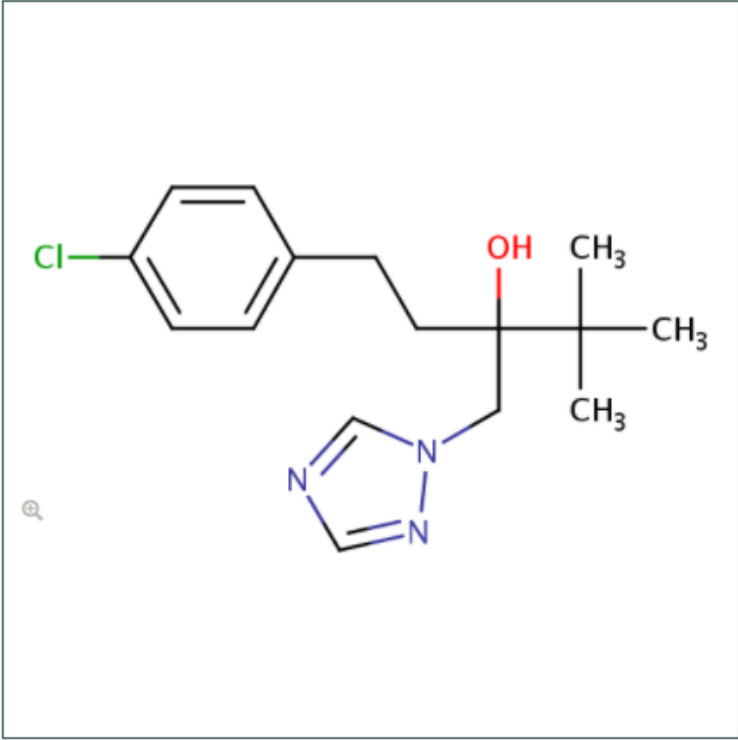
Once the user is satisfied with the number of analogues, clicking on the "Next" button as shown in Figure 16 proceeds to the next step of the workflow.

Figure 16: Proceeding to the Data Gap Analysis step in the workflow

Figure 17: Summary Data Gap Analysis view
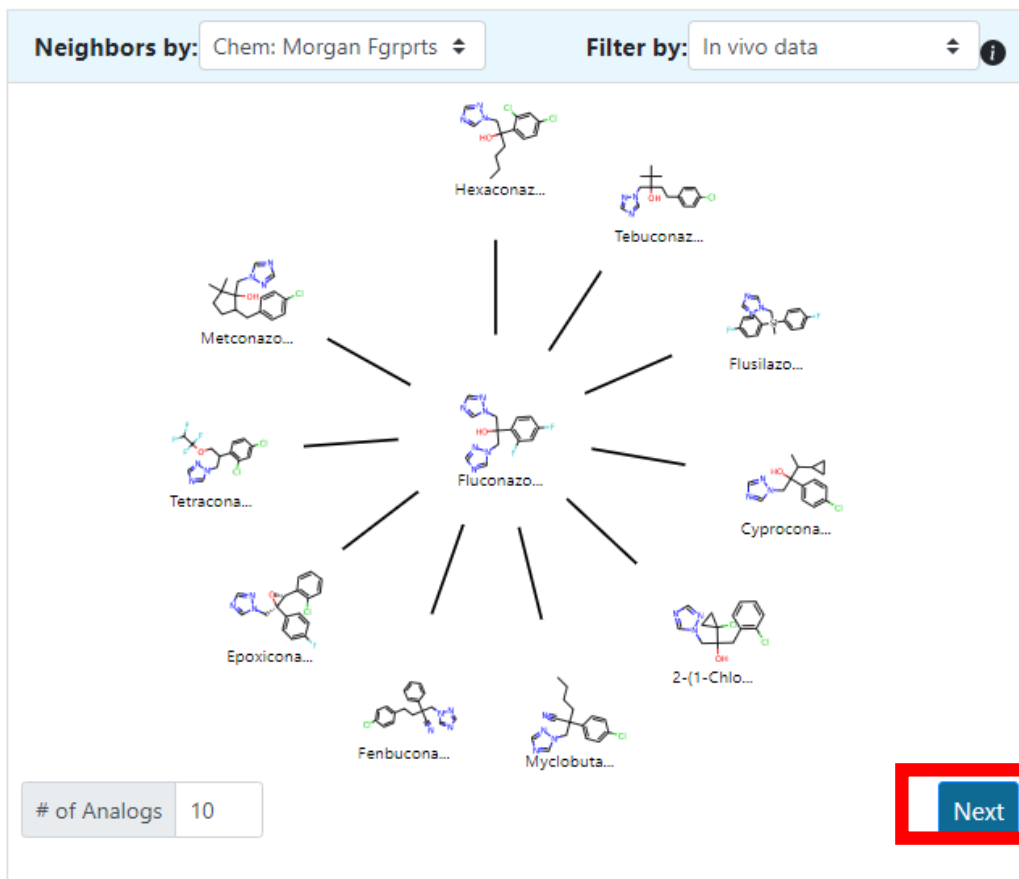
Data gap Analysis

A summary overview of the available data quantity for the target chemical Fluconazole and its source analogues is provided in the second grid (Figure 17). The color density represents a measure of 'data availability' for the target – from light to dark. The number of data records is reflected in the box itself. The data availability is segmented by data type – where bio_txct represents bioactivity data from ToxCast, e.g. chm_ct represents chemistry ToxPrint descriptor information and tox_txrf represents *in vivo* toxicity effect information from ToxRefDB v2.0. The number is an indicator of the number of 'hits' in the fingerprint representation of that data type – i.e. large numbers for bio_txct would be indicative of number of active hitcalls within the spectrum of assays which make up the fingerprint representation whereas number of structural bits would be the equivalent in the chemical fingerprint. In the case of target Fluconazole, there are 'some' bioactivity and chemistry data (based on the colours in the boxes) from which toxicity predictions could be made but there are no *in vivo* toxicity data available in ToxRefDB v2.0.

Figure 17: Summary Data Gap Analysis view
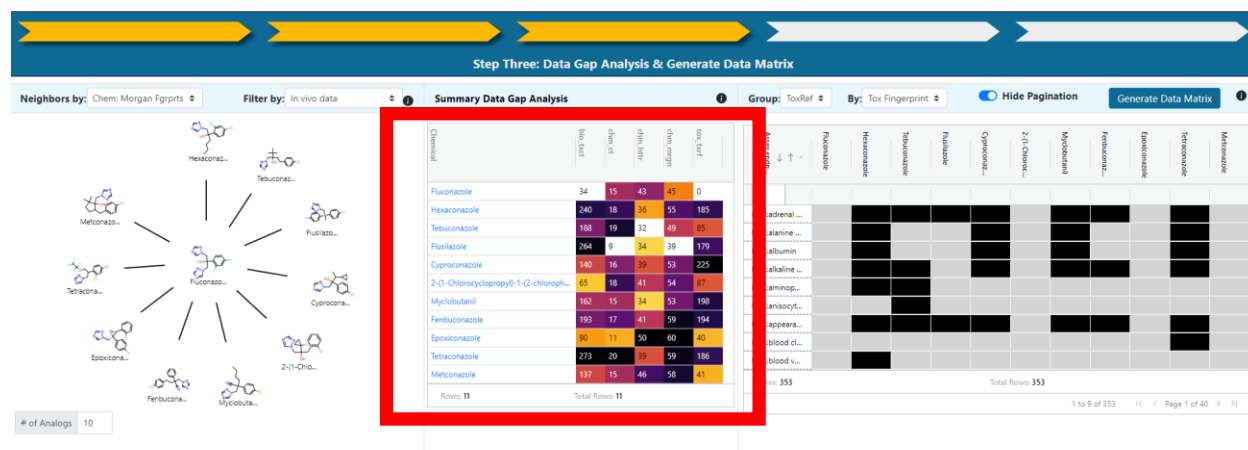
The third grid (Figure 17) represents the available data captured on the basis of the toxicity effects within the ToxRefDB v2.0 studies. Here a box marked in black indicates the availability of information vs lack of any information. Once the user has browsed the matrix to identify what types of data 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 summarises the same information but on the basis of activity score (presence or absence 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 Figure 17 provides 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 17, the Tox fingerprint reflects the toxicity effects within each study type. Since there are over 300 different study type-toxicity effect combinations represented in ToxRefDB v2.0, the user can use the paginate option to scroll through the toxicity effects represented.

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 18. This produces a data matrix view. Across the top are the target chemical, the source analogues ordered by similarity. Each row represents a study type-toxicity effect.

Figure 18: Data matrix view
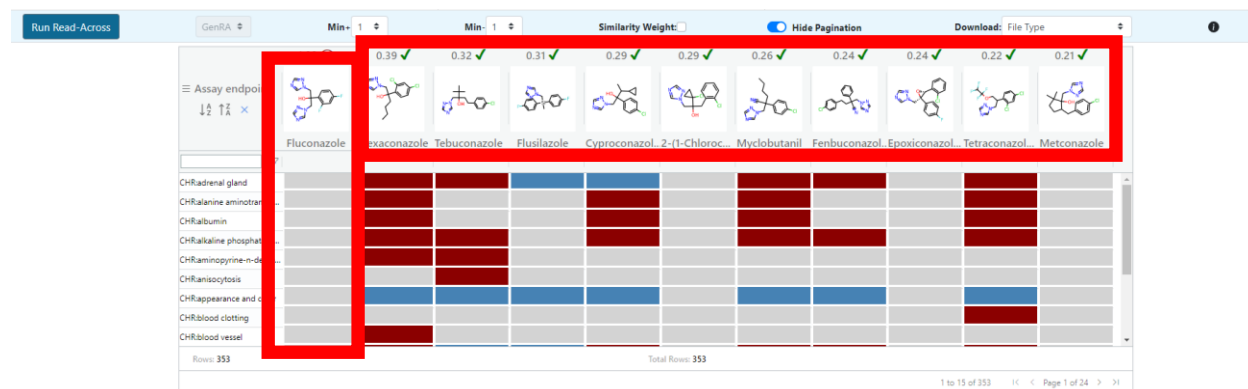
The toxicity effects are populated by red and blue boxes across the analogues representing the presence or absence of toxicity effects in the *in vivo* studies from ToxRefDB v2.0. Blue indicates an absence of effects whereas red represents presence of toxicity effects. Grey boxes indicate an absence 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 more 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 19.

Figure 19: Data gap filling by GenRA
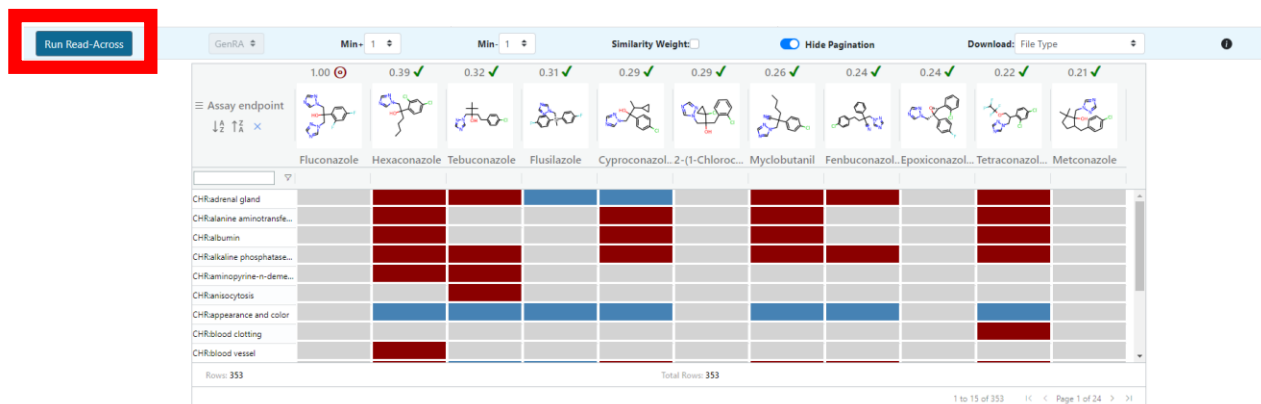
Predictions will by default be run on the screen displayed. The toxicity effects being predicted can also be filtered using the "Filter" box. The predictions can also be modified by specifying thresholds for the number of actives and inactives across the analogues using the Min+ and Min- filters. The 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 20).

Figure 20: De-selecting analogues within the data matrix
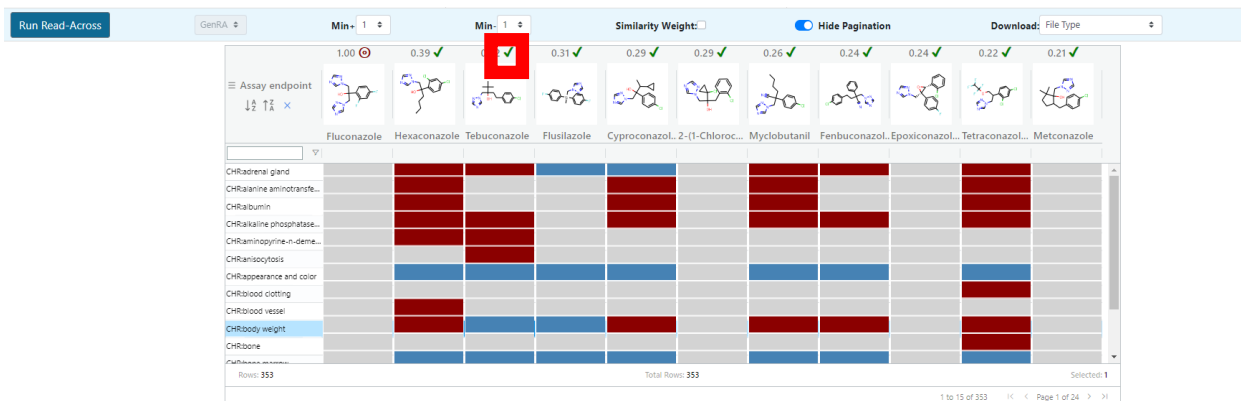
Clicking on "Run Read-across" will update the target information to show the predictions made (Figure 21). The opacity of the predictions reflects the confidence in the prediction made with. A faint colored prediction will denote lower confidence in the prediction.
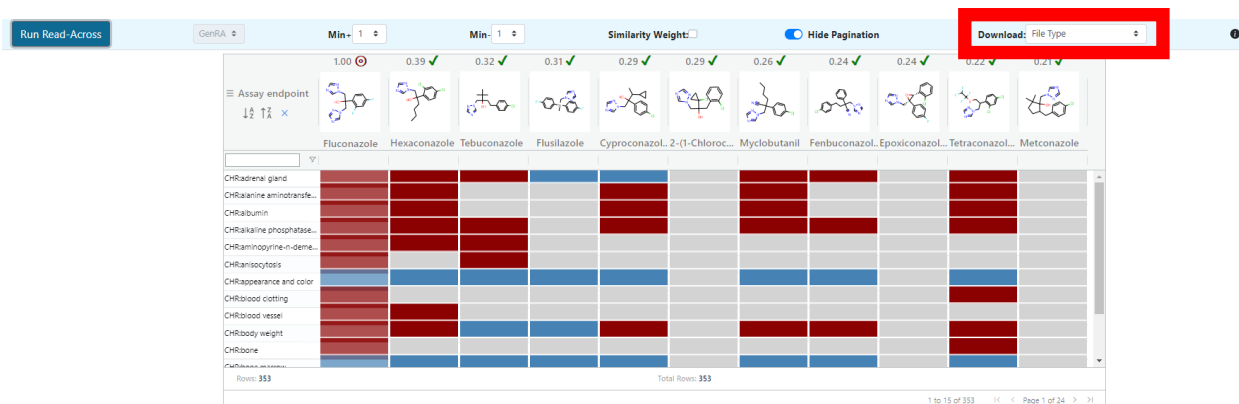
Figure 21: Data gap filling predictions by GenRA
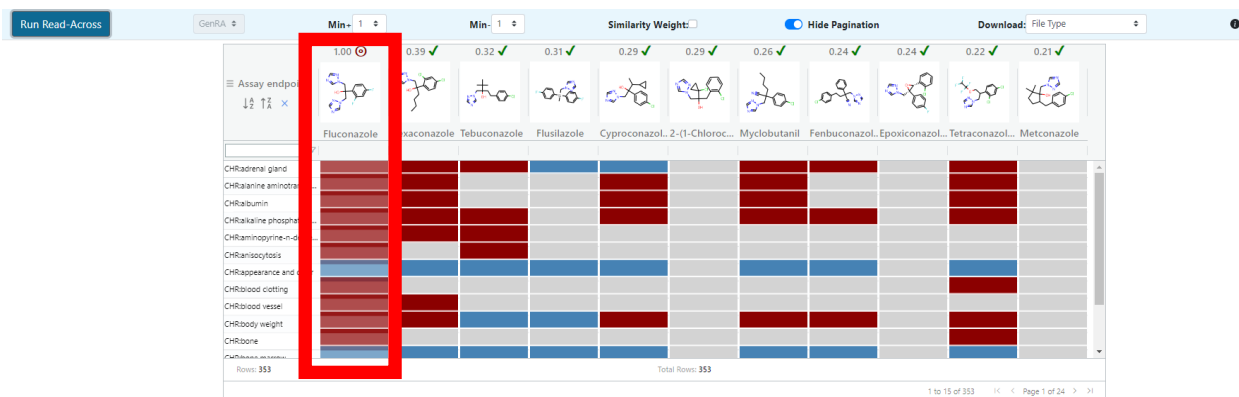
Uncertainty assessment

Predictions made can be exported by clicking on the Download button shown in Figure 21.

Figure 21: 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 downloaded as a CSV file or Excel (xlsx) file. A snapshot of what the download file resembles from Excel is depicted in Figure 22.

Figure 22: Sample prediction output produced in the Excel (xlsx) export file

chem_id	DTXCID10627	DTXCID2014653	DTXCID2014653_u	DTXCID7012113	DTXCID7012113_u	DTXCID704235	DTXCID704235_ur	DTXCID8012601	DTXCID8012601_u	DTXCID1024338	DTXCID1024338_u	DTXCID304315	DTXCID304315_ur
role	target	analog		analog		analog		analog		analog		analog	
preferred name	Fluconazole	Hexaconazole		Tebuconazole		Flusilazole		Cyproconazole		2-(1-Chlorocyclopropyl)-1-(2-chloro		Myclobutanil	
dsstox_sid	DTXSID3020627	DTXSID4034653		DTXSID9032113		DTXSID3024235		DTXSID0032601		DTXSID3044338		DTXSID8024315	
dsstox_cid	DTXCID10627	DTXCID2014653		DTXCID7012113		DTXCID704235		DTXCID8012601		DTXCID1024338		DTXCID304315	
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
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
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	
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	
CHR:appearance and color	GenRA Neg; ACT=None; AUC=0; pval=1	no_data		no_effect		no_effect		no_effect		no_data		no_effect	
CHR:blood clotting	GenRA Pos; ACT=None; AUC=0; pval=1	no_data		no_data		no_data		no_data		no_data		no_data	
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	
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	
CHR:bone marrow	GenRA Neg; ACT=None; AUC=0; pval=1	no_effect		no_effect		no_effect		no_effect		no_data		no_effect	
CHR:brain	GenRA Neg; ACT=None; AUC=0; pval=0.93	no_effect		no_effect		no_effect		no_data		no_data			393.5 mg/kg/day
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	
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_effect		no_data	
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	
CHR:ear	GenRA Neg; ACT=None; AUC=0; pval=1	no_effect		no_effect		no_effect		no_effect		no_data		no_effect	
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
CHR:erythrocyte (rbc) count	GenRA Neg; ACT=None; AUC=0.0; pval=0.84	no_effect		no_data		no_data		no_effect		no_data			40 mg/kg/day
CHR:esophagus	GenRA Neg; ACT=None; AUC=0; pval=1	no_effect		no_effect		no_effect		no_effect		no_effect		no_effect	
CHR:eye	GenRA Neg; ACT=None; AUC=0.0; pval=0.86	no_effect			77.3 mg/kg/day	no_effect		no_effect		no_effect		no_effect	
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
CHR:full gross necropsy	GenRA Neg; ACT=None; AUC=0; pval=1	no_effect		no_effect		no_effect		no_effect		no_effect			40 mg/kg/day
CHR:gamma glutamyl transfer	GenRA Pos; ACT=None; AUC=0; pval=1	no_data		no_data		no_data		no_data		no_data			40 mg/kg/day
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	
CHR:glucose	GenRA Neg; ACT=None; AUC=0.0; pval=0.875	no_effect		no_effect		no_effect		no_effect		no_data		no_effect	
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	
CHR:heart	GenRA Neg; ACT=None; AUC=0.0; pval=0.85	no_effect		no_effect		no_effect		no_effect		no_data		no_effect	
CHR:hematocrit (hct)	GenRA Pos; ACT=None; AUC=0; pval=1	no_data		no_data		no_data		no_data		no_data		no_data	

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.

Questions/Further information

For further information: contact us at genra.support@epa.gov. This email is provided as the Contact link on the main application page.

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.* 8: 34-50.

Helman G, Shah I, Williams AJ, Edwards J, Dunne J, Patlewicz G. 2019. Generalized Read-Across (GenRA): A workflow implemented into the EPA CompTox Chemicals Dashboard. *ALTEX.* 36(3): 462-465. doi: 10.14573/altex.1811292.

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.

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.