EPA United States Environmental Protection Agency

Influence of Transcriptomic Descriptors on the Generalised Read-Across (GenRA) Performance

Tia Tate[1], Grace Patlewicz[1], John Wambaugh[1], Imran Shah[1]

[1]Center for Computational Toxicology and Exposure, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

Conflict of Interest Statement

No conflict of interest declared.

Disclaimer:

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

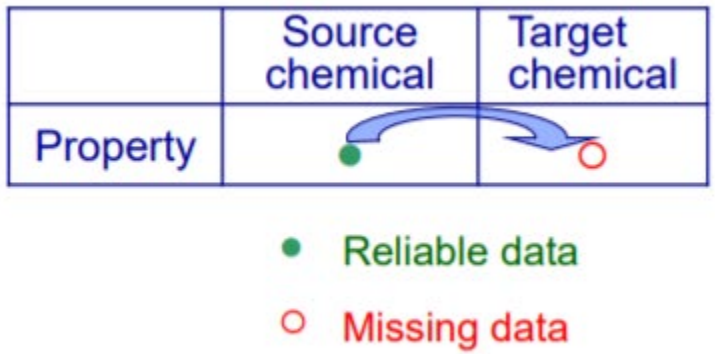
EPA United States Environmental Protection Agency

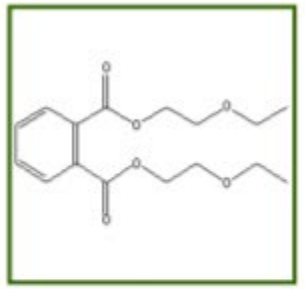
Outline

- Overview of the Generalised Read-Across (GenRA) Approach
- Using GenRA standalone for prediction of toxicity with chemical structure and transcriptomic descriptors
- Evaluation of predictions
- Future work & conclusions

Background & Definitions

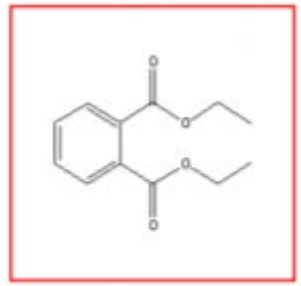
- Read-across is a data gap filling technique utilized to predict the toxicity of a target chemical using toxicity data from source analogues that have similar properties.
  - A target chemical is a chemical which has a data gap that needs to be filled i.e. the subject of the read-across.
  - A source 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.





Known to be harmful

Liver Toxicity??



Predicted to be harmful

Read-Across Tools

Computational Toxicology 3 (2017) 1–18

Contents lists available at ScienceDirect

Computational Toxicology

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

ELSEVIER

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 201 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, structur 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 provid	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 file of prediction report, text file of data, image files of plots e	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.

Generalised Read-Across (GenRA)

- The Generalised Read-Across (GenRA) approach facilitates automated read-across predictions for untested chemicals.
- Aims to make binary and quantitative predictions of toxicity outcomes based on neighboring chemicals characterized by chemical and/or bioactivity descriptors (Shah et al, 2016).
- Current version available on the EPA CompTox Chemicals Dashboard.

Computational Toxicology 8 (2018) 34–50

Contents lists available at ScienceDirect

Computational Toxicology

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

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

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

[a] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA
[b] 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

ARTICLE INFO

Keywords:
Read-across
Generalised Read-Across (GenRA)
Similarity in bioavailability
Physicochemical parameters
Read-across performance

ABSTRACT

Read-across is a useful data gap filling technique used within category and analogue approaches in regulatory hazard and risk assessment. Recently we developed an algorithmic, approach called Generalised Read-Across (GenRA) (Shah et al., 2016) which makes read-across predictions of toxicity effects using a similarity weighted average of source analogues characterised by their chemical and/or bioactivity descriptors. A default GenRA approach (termed baseline GenRA) relies on identifying 10 source analogues relative to a target substance that are structurally similar based on Morgan chemical fingerprints and computing an activity score to estimate presence or absence of *in vivo* toxicity. This current study investigated the impact that similarity in bioavailability plays in altering the local neighbourhood of source analogues as well as read-across performance relative to baseline GenRA using physicochemical property information as a surrogate for bioavailability. Two approaches were evaluated: (1) a filtering approach which restricted structurally related analogues based on their physicochemical properties; and (2) a search expansion approach which included additional analogues based on a combined structural and physicochemical similarity index. Filtering minimally improved performance, and was very dependent on the similarity threshold selected. The search expansion approach performed at least as well as the baseline GenRA, and showed up to a 9% improvement in read-across performance for at least 10 of the 50 organs considered. We summarise the overall impact that physicochemical information plays on GenRA performance, illustrate the improvement for a specific case study substance and describe how to select the most appropriate physicochemical similarity threshold to achieve optimal read-across performance depending on the toxicity effect and chemical of interest. The analyses show that physicochemical property information does result in a modest (up to 9% increase) improvement in structural based read-across predictions.

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.

Computational Toxicology 12 (2019) 100097

Contents lists available at ScienceDirect

Computational Toxicology

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

Transitioning the generalised read-across approach (GenRA) to quantitative predictions: A case study using acute oral toxicity data

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

[a] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA
[b] 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

ARTICLE INFO

Keywords:
Generalized read across (GenRA)
Acute oral toxicity
Quantitative predictions

ABSTRACT

Read-across approaches continue to evolve as does their utility in the field of risk assessment. Previously we presented our generalised read-across (GenRA) approach (Shah et al., 2016), which utilises chemical descriptor and/or *in vitro* bioactivity data to make read-across predictions on the basis of the similarity weighted average of nearest neighbours. The current public version of GenRA predicts 574 apical outcomes as a binary call from repeat dose toxicity studies available in ToxRefDB (Helman et al., 2019). Here we investigated the application of GenRA to quantitative values, specifically using a large dataset of rat oral acute LD50 toxicity data (LD50 values for 7011 discrete chemicals) that had been collected under the auspices of the ICCVAM acute toxicity workgroup (ATWG). GenRA LD50 predictions were made based on the following criteria – chemicals were characterised by Morgan chemical fingerprints with a minimum similarity threshold of 0.5 and a maximum of 10 nearest neighbours over the entire dataset. An R^2 value of 0.61 and RMSE of 0.58 was achieved based on these parameters. Monte Carlo cross validation was then used to estimate confidence in the R^2. Cross validated R^2 values were found to fall in the range of 0.47–0.62. However, when evaluating GenRA locally to clusters of mechanistically or structurally-similar chemicals, average R^2 values improved up to 0.91. GenRA can be extended to make reasonable quantitative predictions of acute oral rodent toxicity with improved performance exhibited for specific local domains.

Regulatory Toxicology and Pharmacology 109 (2019) 104480

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Quantitative prediction of repeat dose toxicity values using GenRA

G. Helman[a,b], G. Patlewicz[b], I. Shah[b,*]

[a] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA
[b] National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

ARTICLE INFO

Keywords:
Computational toxicology
Generalized read-across
Point of departure
Repeat-dose toxicity
Cheminformatics

ABSTRACT

Computational approaches have recently gained popularity in the field of read-across to automatically fill data-gaps for untested chemicals. Previously, we developed the generalized read-across (GenRA) tool, which utilizes *in vitro* bioactivity data in conjunction with chemical descriptor information to derive local validity domains to predict hazards observed in *in vivo* toxicity studies. Here, we modified GenRA to quantitatively predict point of departure (POD) values obtained from US EPA's Toxicity Reference Database (ToxRefDB) version 2.0. To evaluate GenRA predictions, we first aggregated oral Lowest Observed Adverse Effect Levels (LOAEL) for 1,014 chemicals by systemic, developmental, reproductive, and cholinesterase effects. The mean LOAEL values for each chemical were converted to log molar equivalents. Applying GenRA to all chemicals with a minimum Jaccard similarity threshold of 0.05 for Morgan fingerprints and a maximum of 10 nearest neighbors predicted systemic, developmental, reproductive, and cholinesterase inhibition min aggregated LOAEL values with R^2 values of 0.23, 0.22, 0.14, and 0.43, respectively. However, when evaluating GenRA locally to clusters of structurally-similar chemicals (containing 2 to 362 chemicals), average R^2 values for systemic, developmental, reproductive, and cholinesterase LOAEL predictions improved to 0.73, 0.66, 0.60 and 0.79, respectively. Our findings highlight the complexity of the chemical-toxicity landscape and the importance of identifying local domains where GenRA can be used most effectively for predicting PODs.

General Approach

I. Data

- Chemical Data (chm)
  - Structural Descriptors (i.e. Morgan fingerprints)
- Bioactivity Data (bio) (i.e. bioactivity assays)
- Toxicity Outcomes (tox) (ToxRefDB)

II. Generate Local Neighborhoods

- Group chemicals using a similarity-weighted activity score of nearest neighbors.
  - Similarity calculated using Jaccard distance.

$$\gamma_i = \frac{\sum_j^k s_{ij} x_j}{\sum_j^k s_{ij}}$$

III. GenRA

- Evaluation of the performance of chm, bio, and hybrid descriptors for the prediction of toxicity outcomes in local neighborhoods .

EPA United States Environmental Protection Agency

Current Application

- Previously, high throughput screening bioactivity data were collected from ToxCast.
- This study investigates the impact of biological similarities (as characterized by transcriptomic data) on local neighborhood formation and overall read-across performance in qualitatively predicting hazard based on toxicological study data summarized in US EPA ToxRefDB v2.0.
- We expanded on the previous approach with an updated data set composed of high throughput transcriptomics biological data from HepaRG™ cells treated with 8 concentrations across 1060 ToxCast chemicals for 93 transcripts.

EPA United States Environmental Protection Agency

Current Application

I. Data

- Chemical Data (chm)
  - Structural Descriptors (i.e. Morgan fingerprints)
- Bioactivity Data (bio)
  - (i.e. bioactivity assays)
- Toxicity Outcomes (tox) (ToxRefDBv2)
- Chemical Clusters
  - Shah el al (2016)

II. Evaluation of Optimal Number of Nearest Neighbors and Similarity Metric

- scikit-learn grid search 5-fold cross validation

III. Generate Local Neighborhoods

- Group chemicals using a similarity-weighted activity score of nearest neighbors.
  - Similarity calculated by:
    - Jaccard
    - Manhattan
    - Euclidean

IV. GenRA

- Global performance evaluation of chm, bio and hybrid descriptors in the prediction toxicity endpoints using area under the ROC curve (AUC).
- Local performance evaluation of chm, bio, and hybrid descriptors in the prediction of toxicity endpoint using predefined chemical cluster using area under the ROC curve (AUC).

GenRA-py

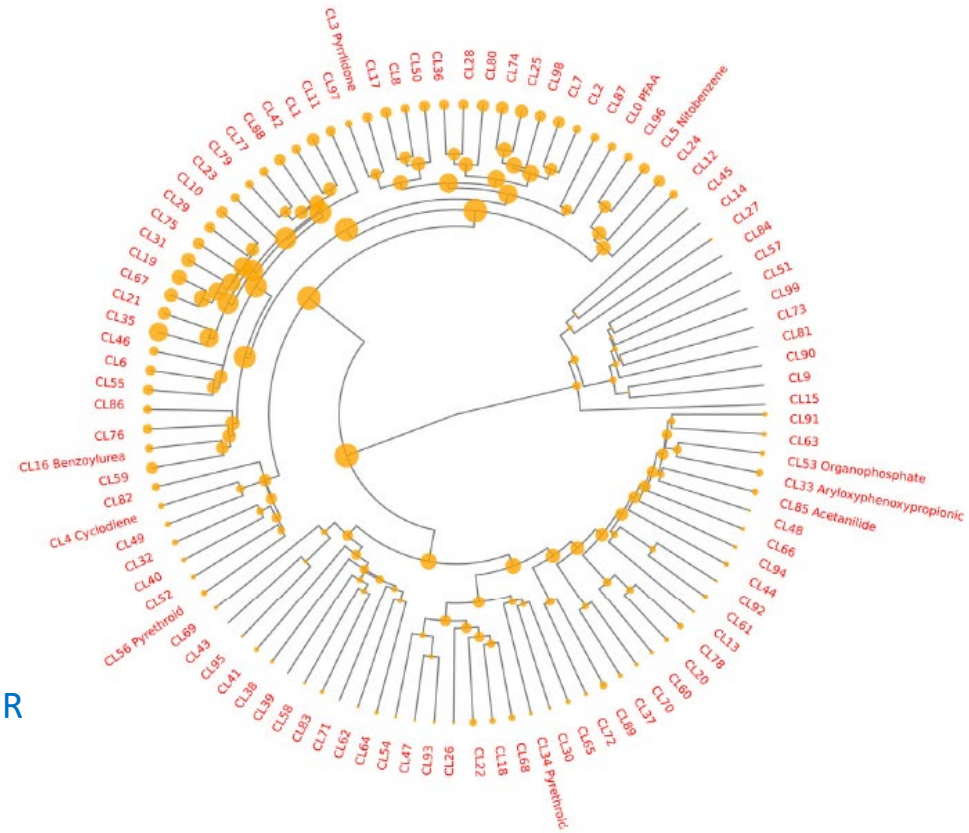
Data

- **Chemical Data (C)**
  - Morgan Chemical Fingerprints (mrgn)
  - Torsion Topological Fingerprints (tptr)
  - ToxPrints (toxp)

- **Chemical Clusters**
  - Identified in Shah et al, 2016

- **Biological Data (B)**
  - HepaRG™ LTEA (Life Technologies/Expression Analysis) Assay
    Wambaugh *et al*, 2020
    - LTEA assay results analyzed with the ToxCast pipeline package in R (tcpl) for curve fitting
      - Assay level hit call data
      - Gene level hit call data

- **Toxicity Data**
  - ToxRefDBv2.0 negative (0) and positive(1) toxicity endpoints for several study types:
    - Chronic (chr)
    - Subchronic (sub)
    - Acute (acu)
    - Subacute (sac)
    - Developmental (dev)
    - Multigeneration reproductive (mgr)
    - Reproductive (rep)
    - Developmental neurotoxicity (dnt)
    - Neurological (neu)
    - Other (oth)

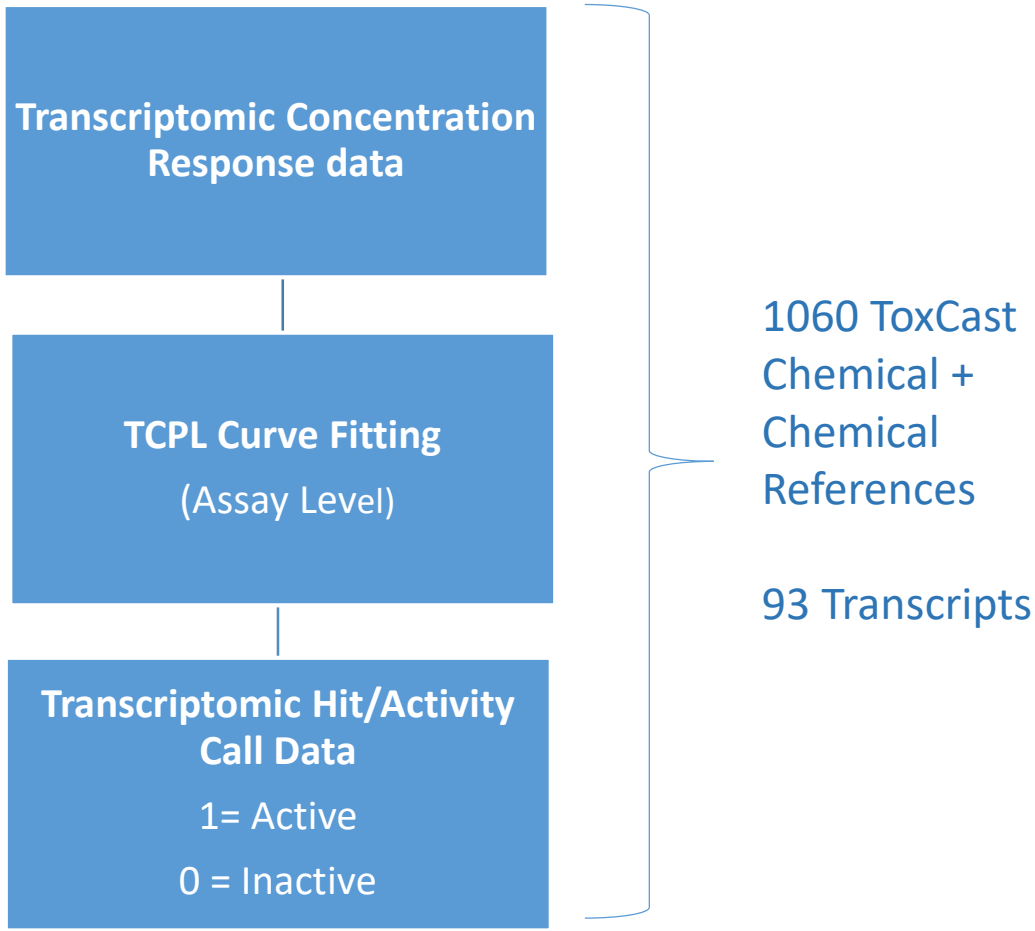


I. Shah et al. / Regulatory Toxicology and Pharmacology 79 (2016) 12–24 17

Fig. 1. Clustering chemicals by structural similarity. The dendrogram shows the results of hierarchical agglomerative clustering of the centroids of all 98 clusters (see Methods). Each leaf node in the tree is a cluster where the number of chemicals in the cluster is proportional to the size of the circle. Some illustrative examples of the predominant chemical classes in clusters are labeled.

HepaRG™ Data

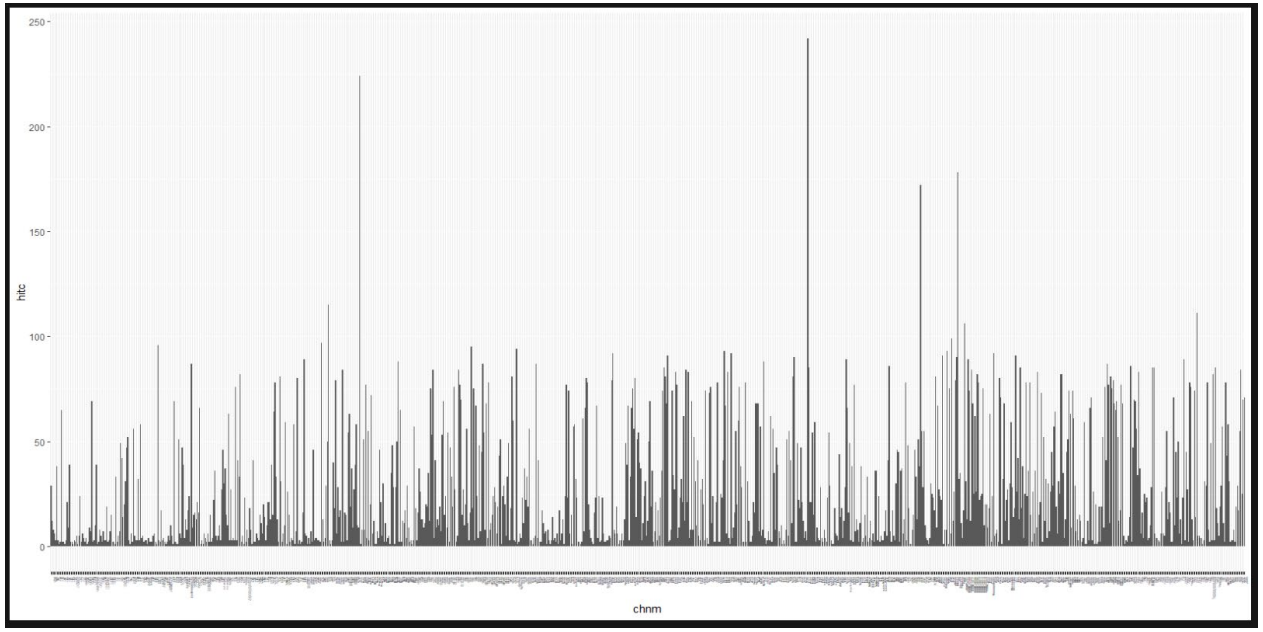
- Treated with 8 concentrations of 1,060 chemicals for 24 hours and the expression of 93 transcripts was measured using quantitative reverse transcription polymerase chain reaction (qRT-PCR).
  - Transcripts measure the expression of genes involved in nuclear receptor activation, xenobiotic metabolism, cellular stress, cell cycle progression, and apoptosis.
- Concentration-response data for the 93 transcripts were analyzed with the ToxCast analysis pipeline package in R (tcpl) for curve fitting.
- The hit-call for each chemical and transcript was assigned a binary active (1) or inactive (0) value based on tcpl level 5 data.
- The transcriptomic data for each chemical was represented using the hit calls in two ways.
  - Vector of binary hit-calls for the 95 genes (termed gene)
  - Vector of binary hit- calls with 190 directional activities of 95 genes (termed assay).



HepaRG LTEA Exploratory Data Analysis

Level 5 Hit Call Data	
#Chems per Assay	1084 chems and controls
#Assays/Genes	189/95
Mean	25.68
SD	29.18
Median	13.00
Min	1.0
Max	242.0

Level 5 Hits Per Chemical



EPA United States Environmental Protection Agency

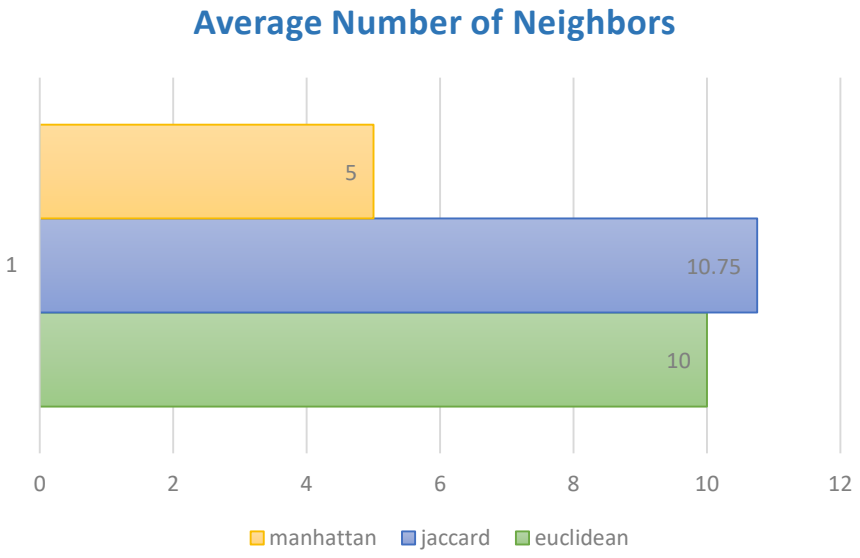
Descriptor Descriptions

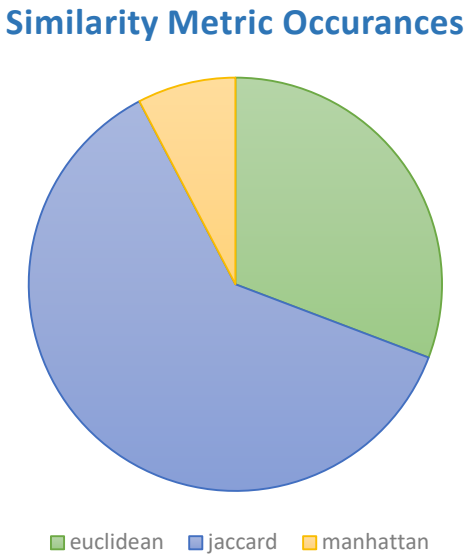
Descriptor Type	Descriptor name	# of Chemicals	# of Descriptors
Chemical(C)	morgan(mrgn)	1017	2048
	torsion (tptr)	1017	2048
	toxprints (toxp)	1017	729
	CA (all chemical, mrgn, tptr, toxp)	1017	4825
Biological (B)	gene	1065	95
	assay	1065	189
Hybrid (CB)	Morgan/gene (mg)	1017	2143
	Morgan/assay (ma)	1017	2237
	Torsion/gene (ttg)	1017	2143
	Torsion/assay (tta)	1017	2237
	ToxPrints/gene (txg)	1017	824
	ToxPrints/assay (txa)	1017	918
	CB (all chemical and Biological, mrgn, tptr, toxp, gene, assay)	1017	5109

Toxicity Data:	# of Chem	# of Study/effects	Study types /Endpoints
ALL	935	922	neu, sub, rep, chr, dnt, sac, mgr, dev, acu, oth
Liver	935	9	chr_liver, dev_liver, dnt_liver, mgr_liver, neu_liver, oth_liver, rep_liver, sac_liver, sub_liver

Performance Tuning

- Conducted 5-fold grid search cross validation with ROC AUC scoring to determine optimal number of neighbors (range 1-15) and distance/similarity metric (Euclidean, Jaccard, Manhattan) for all descriptor types.

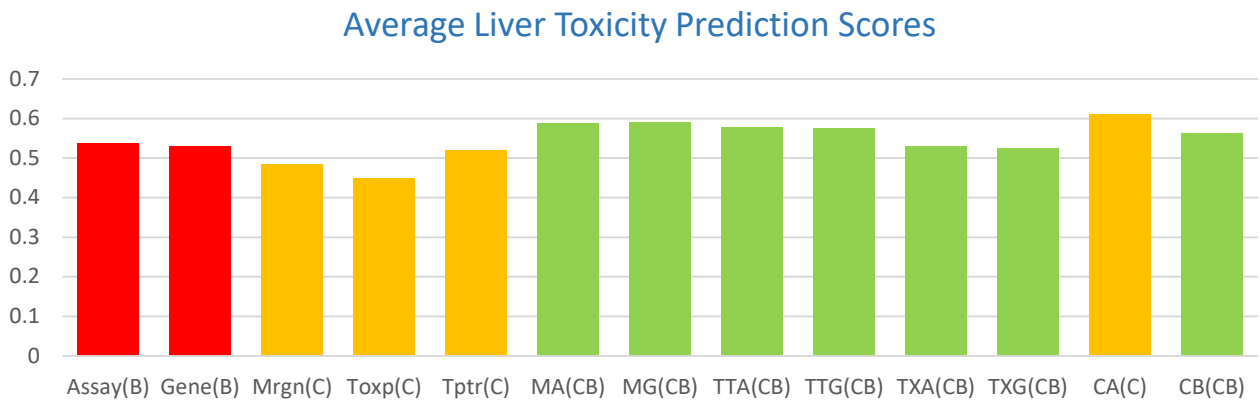


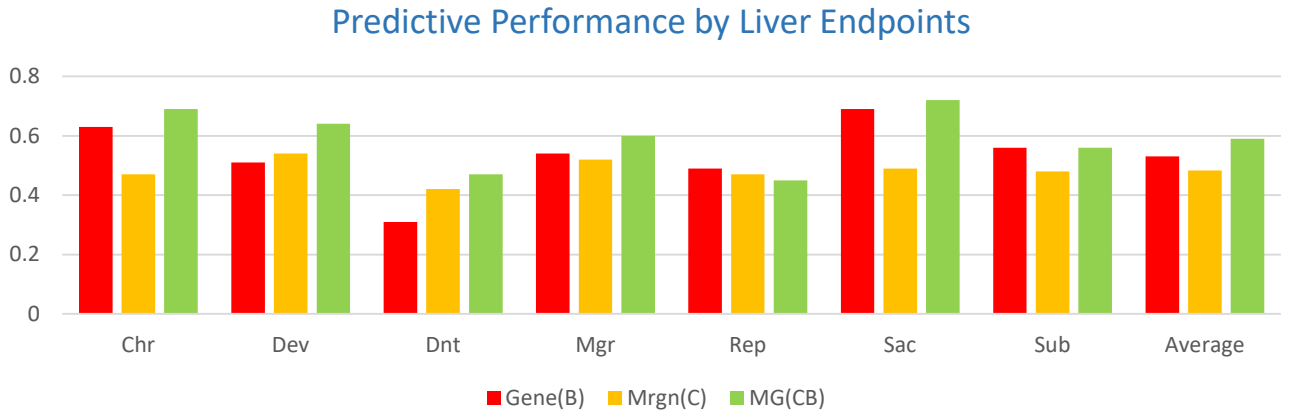


Liver Effect	Descriptor Type	Descriptor Name	AUC	Metric	Number of Neighbors
Chr_liver	Chm	tptr	0.6303	euclidean	9
	Chm	mrgn	0.64549	jaccard	8
	Chm	toxp	0.61379	jaccard	7
	Bio	gene	0.648847	euclidean	14
	Bio	assay	0.6632	euclidean	11
	CB	mrgn/assay	0.6883	jaccard	13
	CB	toxp/gene	0.7044	jaccard	10
	CB	tptr/gene	0.6818	euclidean	6
	CB	(CB) all	0.6999	jaccard	14
	Chm	(CA) all	0.6702	jaccard	10
	CB	mrgn/gene	0.7049	jaccard	10
	CB	toxp/assay	0.6992	jaccard	14
	CB	tptr/assay	0.6721	manhattan	5

Evaluating Overall Global Performance for Prediction of Liver Toxicity Endpoints

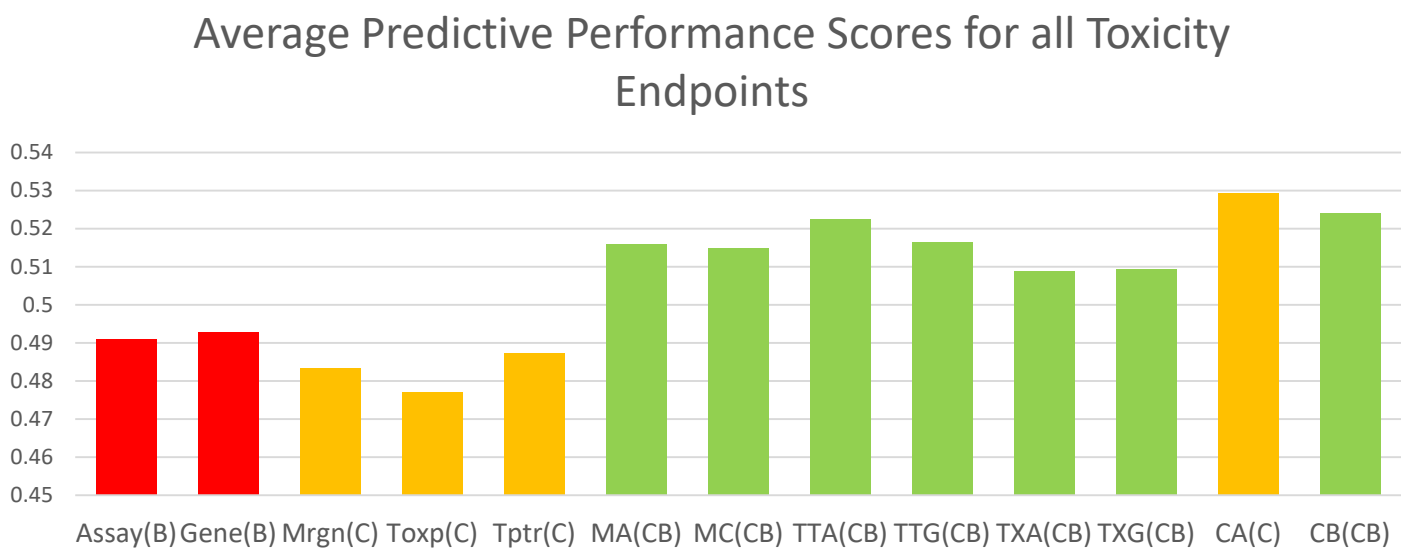
- Biological descriptors outperformed the singular chemical descriptors.
  - 10% increase in predictive performance.
- Hybrid descriptors generated an overall 16% increase in predictive performance in comparison to singular chemical descriptors and a 6% increase in comparison to biological.
- The all chemical descriptor combination outperformed all other descriptors and combinations for predicting liver toxicity
  - 9% increase over the hybrid descriptors.
  - 15% increase over the biological descriptors.
  - 27% increase over the individual chemical descriptors.





Evaluating Overall Global Performance for Prediction of All Toxicity Endpoints

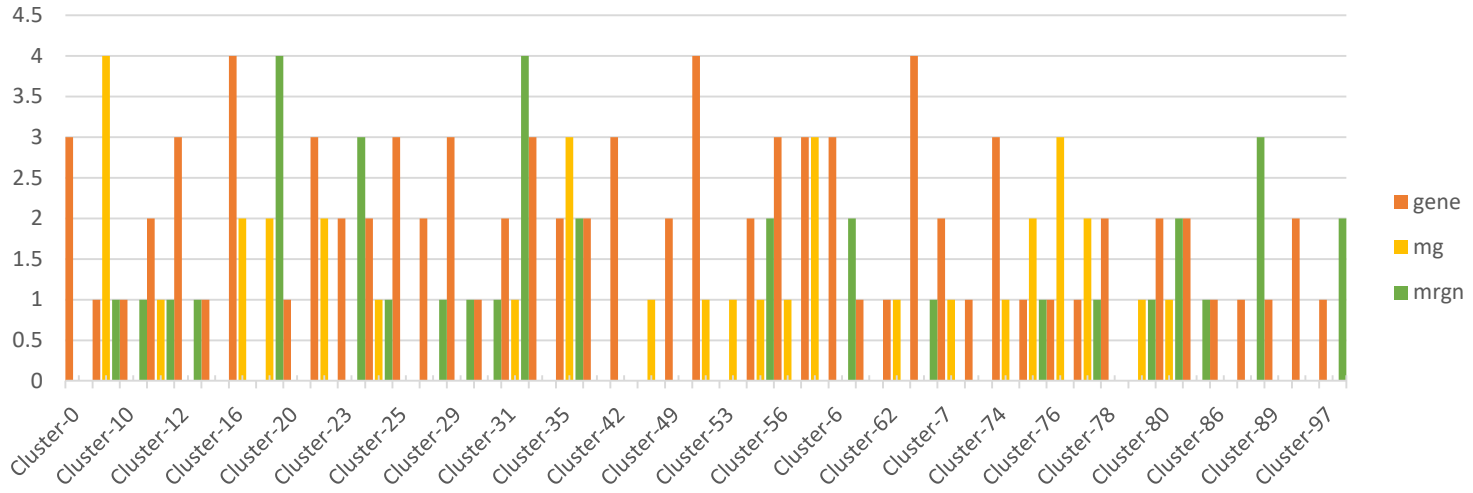
- Toxicity endpoints were aggregated by study type.
- Overall performance score for each study type was calculated.
- Hybrid descriptors consistently outperformed the individual descriptors for the prediction of all toxicity endpoints.
- Chemical hybrid descriptors consisting of all chemical structure fingerprints had the best predictive performance overall.

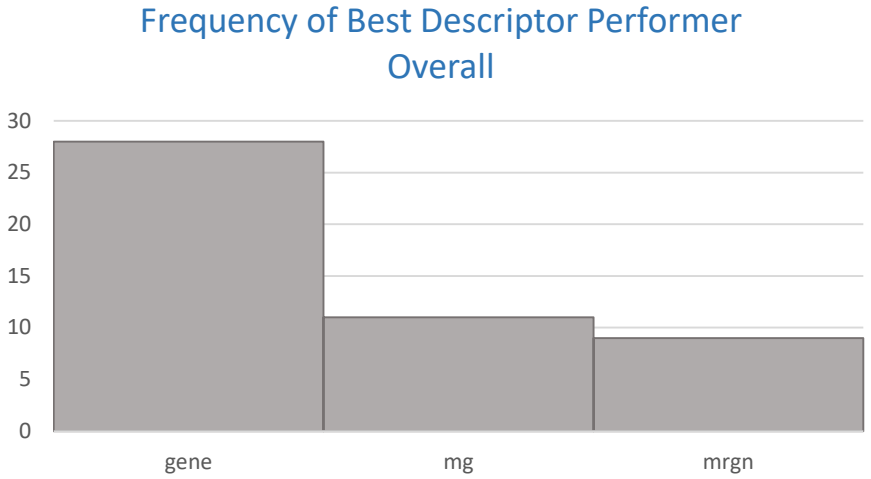


Study	Gene(B)	Mrgn(C)	MG(CB)	B >= 70	C >=70	CB >=70	B > CB | C	C> B |CB	CB > B |C
Chr (364)	0.51 | 0.07	0.50 | 0.07	0.54 | 0.09	2 | 1.2%	2 | 1.2%	7 | 4.2%	41 | 25%	40 | 24%	86 | 51%
Dev (43)	0.49 | 0.07	0.49 | 0.08	0.50 | 0.09	2 | 1.7%	4 | 3.4%	3 | 2.6%	29 | 25 %	45 | 38%	43 | 37%
Dnt (61)	0.44 | 0.12	0.42 | 0.11	0.48 | 0.13	1 | 1.5 %	2 | 3.0	4 | 6.1%	12 | 18%	20 | 30%	34 | 52%
Mgr (201)	0.48 | 0.09	0.48 | 0.09	0.50 | 0.11	3 | 2.2%	3 | 2.2%	8 | 6.0%	39 | 29%	40 | 30%	55 | 41%
Rep (51)	0.45 | 0.12	0.45 | 0.15	0.45 | 0.14	1 | 1.5%	6 | 9.0%	4 | 6.0%	23 | 34%	20 | 30%	24 | 36%
Sac (99)	0.49 | 0.11	0.46 | 0.12	0.50 | 0.12	6 | 4.7%	5 | 3.9%	8 | 6.3%	41 | 32%	45 | 35%	41 | 32%
Sub (315)	0.50 | 0.07	0.49 | 0.08	0.54 | 0.10	3 | 1.8%	5 | 3.0%	12 | 7.1%	36 | 21%	40 | 24%	93 | 55%
ALL	0.49 | 0.09	0.48 | 0.10	0.51 | 0.11	18 | 2.26%	27 | 3.39%	46 | 5.9%	221 | 26%	250 | 30%	376 | 44%

Evaluating Local Performance

- Explored performance of the basis of individual clusters
- Filtered clusters consisting of 2 or more positive and negative endpoints
- Identified clusters where each individual descriptors outperformed the others

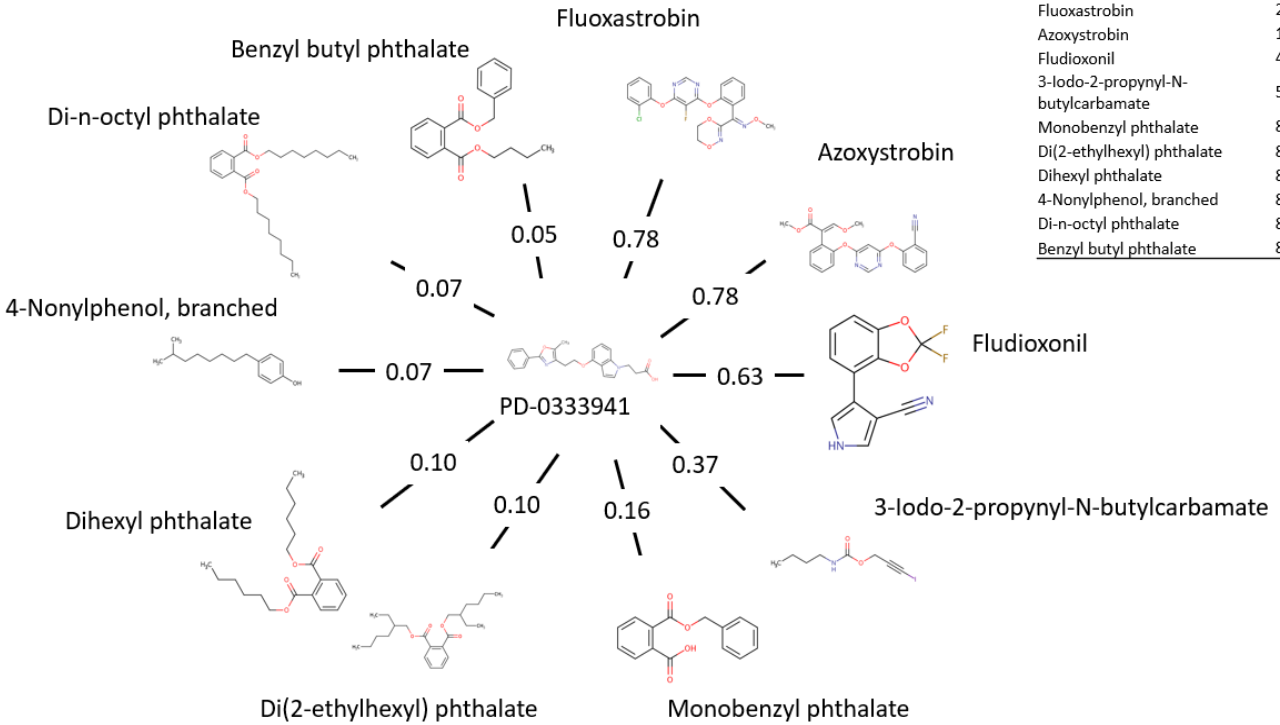




Custer	Study	Gene(B)	Mrgn(C)	MG(CB)	B>=70	C>=70	CB>=70	B > CB | C	C> B |CB	CB > B |C
67	Chr(3)	0.61 | 0.19	0.44 | 0.13	0.57 | 0.22	1 | 33.33%	0 | 0	1 | 33.33%	1 | 33.33%	0 | 0	2 | 66.67%
	Dev(3)	0.6 | 0.3	0.62 |I 0.38	0.36 | 0.23	2 | 66.67%	1 | 33.33%	0 | 0	3 | 100%	0 | 0	0 | 0
	Mgr(2)	0.75 | 0.35	0.48 | 0.26	0.24 | 0.22	1 | 50%	0 | 0	0 | 0	1 | 50%	1 | 50%	0 | 0
	Sac(2)	1 | 0	1 | 0	0.5 | 0	2 | 100%	2 | 100%	0 | 0	1 | 50%	1 | 50%	0 | 0
	Sub(3)	0.44 | 0.19	0.29 | 0.16	0.24 | 0.13	0 | 0	0 | 0	0 | 0	1 | 33.33%	2 | 66.67%	0 | 0
	All(13)	0.65 | 0.27	0.54 | 0.30	0.38 | 0.21	6 | 46.15%	3 | 23. 08%	1 | 7.69%	7 | 53.84%	4 | 30.77%	2 | 15.38%

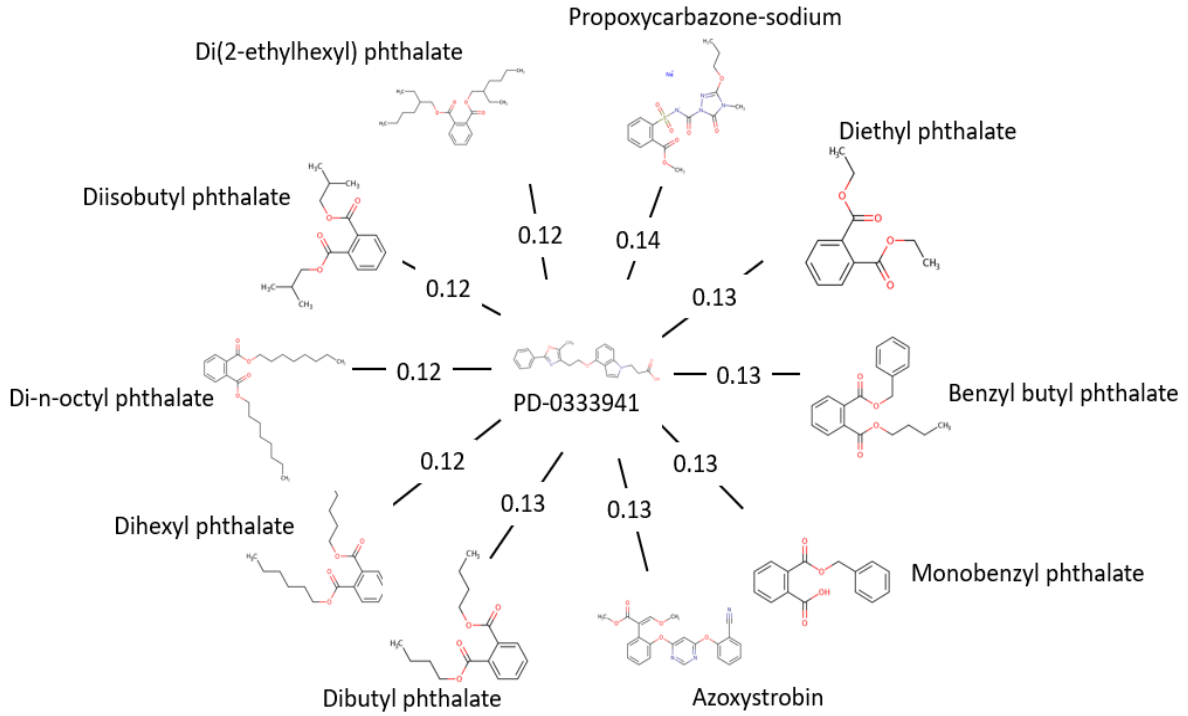
Example Nearest Neighborhood Prediction for Target Chemical in Cluster-80

- Target Chemical: PD-0333941
- Calculation of similarity between target chemical and other chemicals in a predefined chemical cluster based on Jaccard similarity of gene descriptors and Morgan chemical structure descriptors.



PREFERRED_NAME	0	1
PD-0333941	23	72
Fluoxastrobin	21	74
Azoxystrobin	16	79
Fludioxonil	45	50
3-Iodo-2-propynyl-N-butylcarbamate	59	36
Monobenzyl phthalate	82	13
Di(2-ethylhexyl) phthalate	87	8
Dihexyl phthalate	87	8
4-Nonylphenol, branched	89	6
Di-n-octyl phthalate	88	7
Benzyl butyl phthalate	88	7

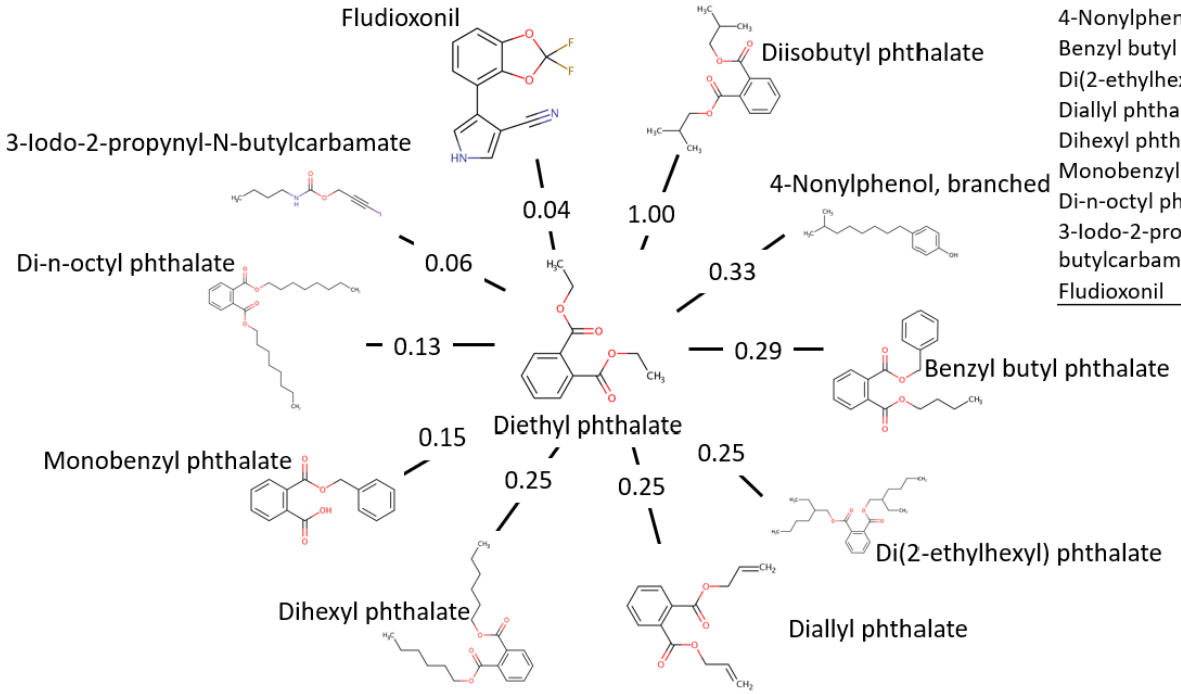
Calculated with Gene Descriptors



Calculated with Morgan Chemical Descriptors

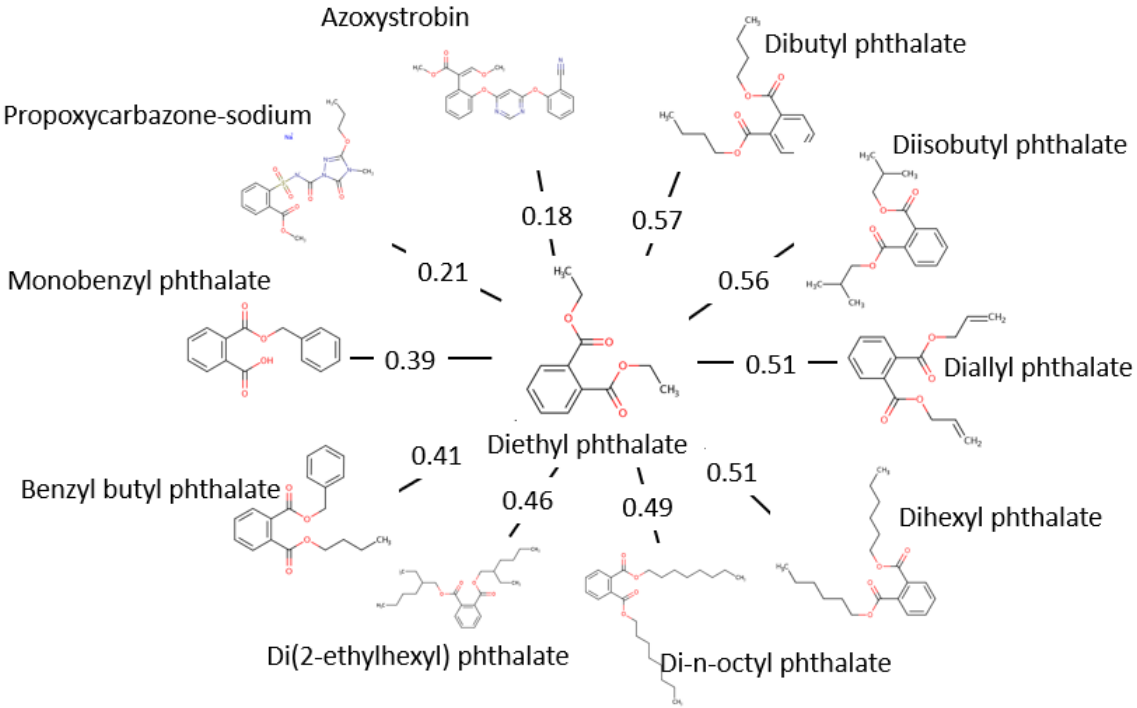
Example Nearest Neighborhood Prediction for Target Chemical in Cluster-80

- Target Chemical: Diethyl phthalate
- Calculation of similarity between target chemical and other chemicals in a predefined chemical cluster based on Jaccard similarity of gene descriptors and Morgan chemical structure descriptors.



PREFERRED_NAME	0	1
Diethyl phthalate	93	2
Diisobutyl phthalate	93	2
4-Nonylphenol, branched	89	6
Benzyl butyl phthalate	88	7
Di(2-ethylhexyl) phthalate	87	8
Diallyl phthalate	92	3
Dihexyl phthalate	87	8
Monobenzyl phthalate	82	13
Di-n-octyl phthalate	88	7
3-Iodo-2-propynyl-N-butylcarbamate	59	36
Fludioxonil	45	50

Calculated with Gene Descriptors



Calculated with Morgan Chemical Descriptors

EPA United States Environmental Protection Agency

Summary

- Chemical structure combination (composed of mrgn, tptr, and toxp) resulted in the best global performance on average for all toxicity endpoints.

- However, an overall increase in read-across performance was noted for various toxicity endpoints when using either transcriptomic and hybrid fingerprints over baseline (mrgn chemical fingerprints).
  - For liver endpoints:
    - Transcriptomic fingerprints resulted in a 10% improvement in performance.
    - Hybrid resulted in a 16% improvement in performance.

- Local predictive performance of various toxicity endpoints across the diverse chemical clusters varied between the diverse set of descriptors.
  - In general, biological descriptors more frequently performed the best across various chemical clusters.

EPA United States Environmental Protection Agency

Future Work and Conclusions

- GenRA was previously shown to predict toxicity using previous HTS of Toxcast compounds but now shown to be applicable on HTTr datasets.
- Here we were able to show that biological descriptors alone or combined with chemical information offer significant benefit in predicting *in vivo* toxicity outcomes on both a ‘global’ and ‘local’ level.
- Future efforts will focus on expanding to diverse/larger transcriptomic data both binary and quantitative.