



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

- Overview of the Generalised Read-across (GenRA) approach
- Using GenRA to predict LD50 from rodent oral acute toxicity studies
- Evaluation of predictions
- Summary Remarks
- Acknowledgements

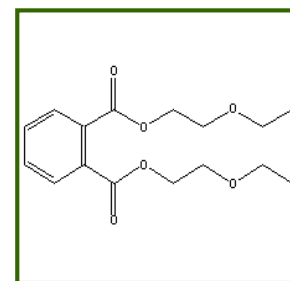
Definitions: Read-across

- Read-across describes the method of filling a data gap whereby a chemical with existing data values is used to make a prediction for a 'similar' chemical.
- A target chemical is a chemical which has a data gap that needs to be filled i.e. the subject of the read-across.
- A source analogue is a chemical that has been identified as an appropriate chemical for use in a read-across based on similarity to the target chemical and existence of relevant data.

	Source chemical	Target chemical
Property		

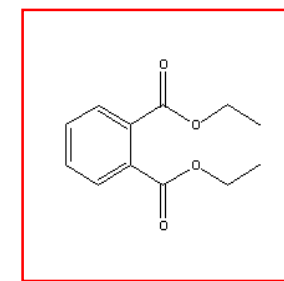
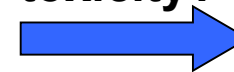
● Reliable data

○ Missing data



**Known to be
harmful**

**Acute
toxicity?**



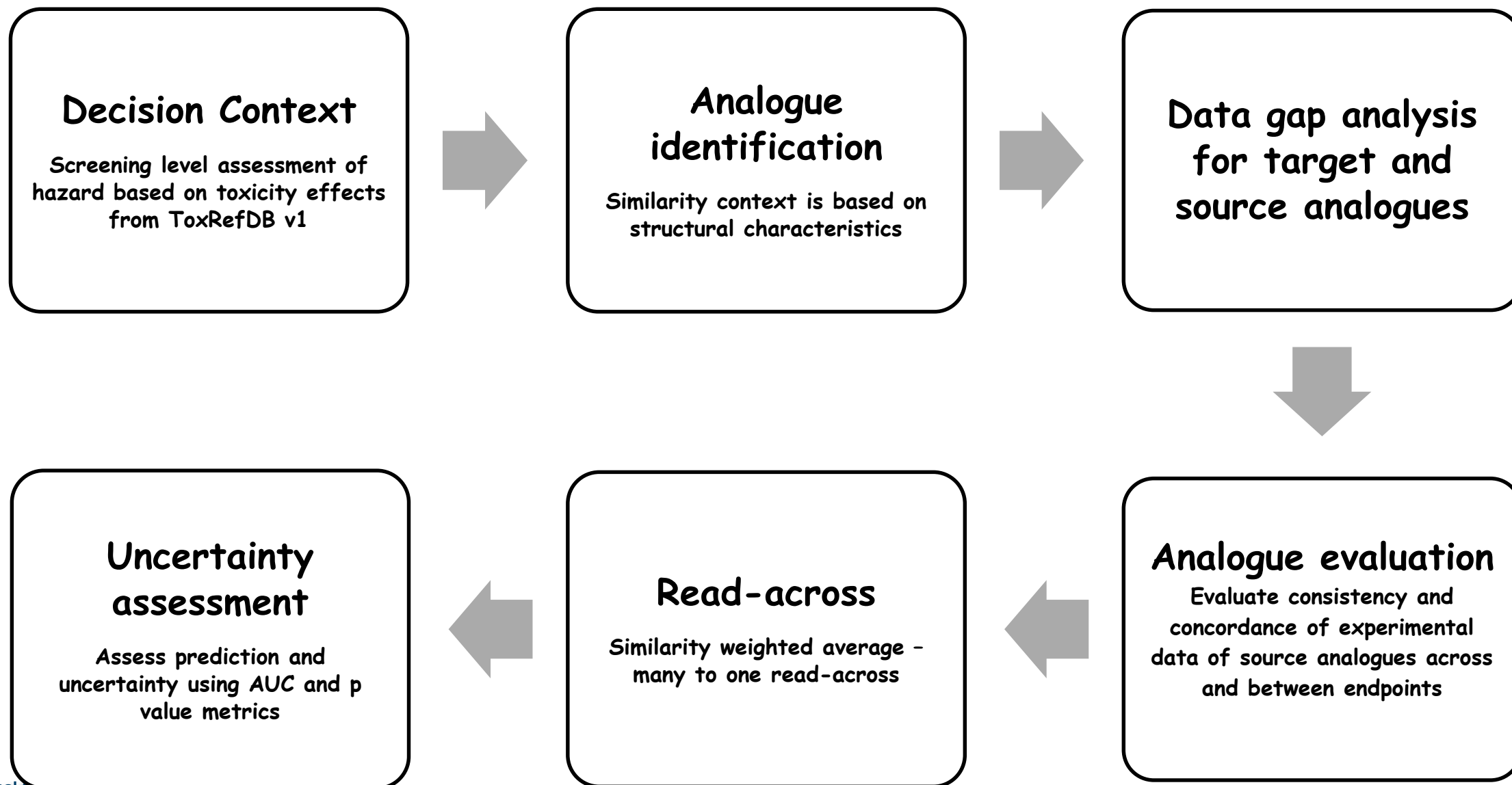
**Predicted to be
harmful**

GenRA (Generalised Read-Across)

- Predicting toxicity as a similarity-weighted activity of nearest neighbours based on chemistry and bioactivity descriptors (Shah et al, 2016)
- Generalised version of the Chemical-Biological Read-Across (CBRA) developed by Low et al (2013)
- Goal: To establish an objective performance baseline for read-across and quantify the uncertainty in the predictions made

Jaccard similarity:

Read-across workflow in GenRA



• Integrat

Short Communication

Generalized Read-Across (GenRA): A workflow implemented into the EPA CompTox Chemicals Dashboard

George Helman^{1,2}, Imran Shah², Antony J. Williams², Jeff Edwards², Jeremy Dunne² and Grace Patlewicz^{2}*

¹Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA; ²National Center for Computational Toxicology (NCCT), Office of Research and Development, US Environmental Protection Agency, Research Triangle Park (RTP), NC, USA

Abstract

Generalized Read-Across (GenRA) is a data driven approach which makes read-across predictions on the basis of a similarity weighted activity of source analogues (nearest neighbors). GenRA has been described in more detail in the literature (Shah et al., 2016; Helman et al., 2018). Here we present its implementation within the EPA's CompTox Chemicals Dashboard to provide public access to a GenRA module structured as a read-across workflow. GenRA assists researchers in identifying source analogues, evaluating their validity and making predictions of *in vivo* toxicity effects for a target substance. Predictions are presented as binary outcomes reflecting presence or absence of toxicity together with quantitative measures of uncertainty. The approach allows users to identify analogues in different ways, quickly assess the availability of relevant *in vivo* data for those analogues and visualize these in a data matrix to evaluate the consistency and concordance of the available experimental data for those analogues before making a GenRA prediction. Predictions can be exported into a tab-separated value (TSV) or Excel file for additional review and analysis (e.g., doses of analogues associated with production of toxic effects). GenRA offers a new capability of making reproducible read-across predictions in an easy-to use-interface.

Refinements to the GenRA approach

- Transitioning GenRA from binary predictions to quantitative predictions
- Investigated extending GenRA using the acute oral rat systemic toxicity data collected as part of the ICCVAM Acute toxicity workgroup
- NICEATM-NCCT effort to collate a large dataset of acute oral toxicity to evaluate the performance of existing predictive models and investigate the feasibility of developing new models

Acute toxicity: Dataset creation

Database Resource	Rows of Data (number of LD50 values)	Unique CAS
ECHA (ChemProp)	5533	2136
JRC AcutoxBase	637	138
NLM HSDB	4082	2238
OECD (eChemPortal)	10206	2314
PAI (NICEATM)	364	293
TEST (NLM ChemIDplus)	13689	13545

Rat oral LD50s:
16,297 chemicals total
34,508 LD50 values

Require unique LD50 values
with mg/kg units

15,688 chemicals total
21,200 LD50 values

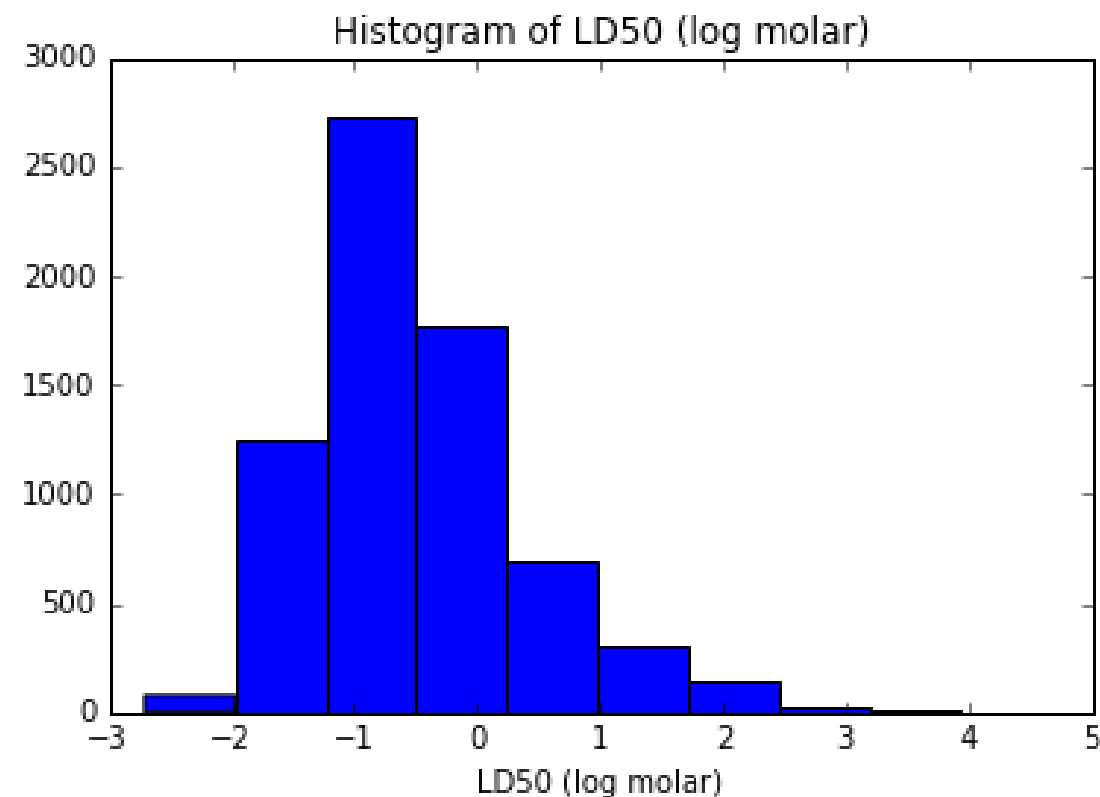
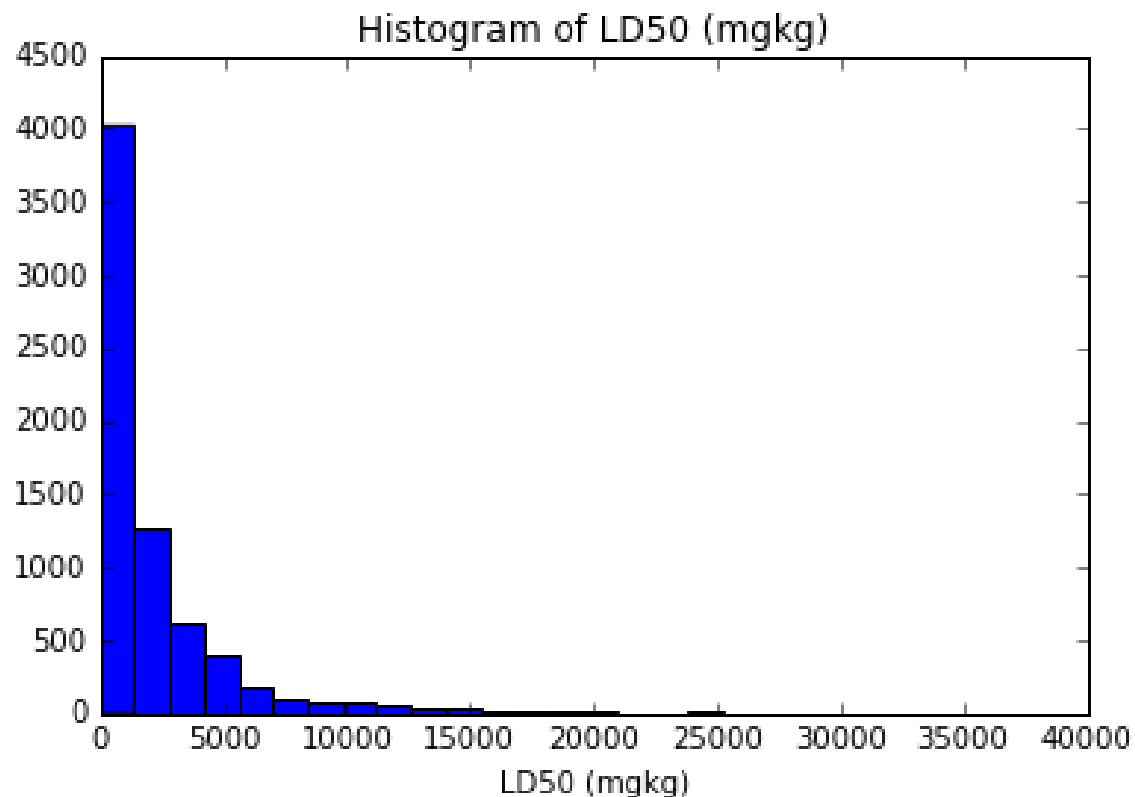
Preprocessing for modelling

11,992 chemicals
16,173 LD50 values

Karmaus et al, 2018; Kleinstreuer et al., 2018

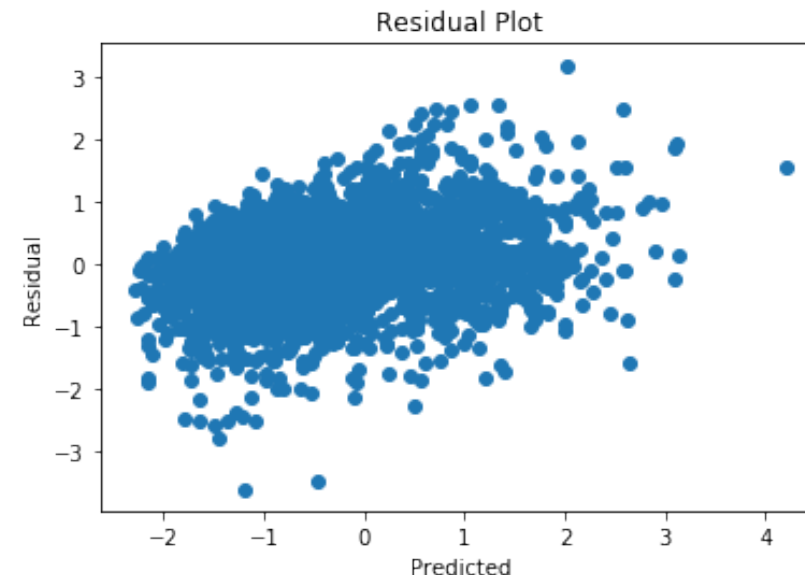
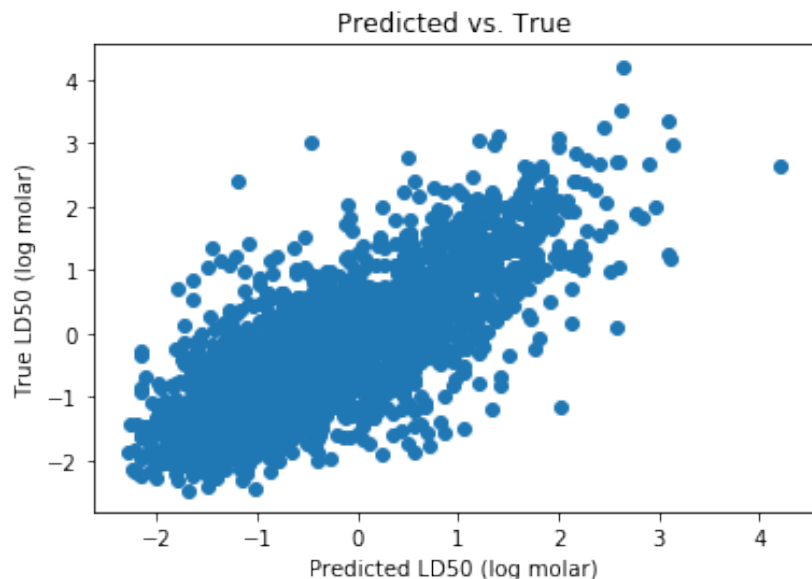
Exploratory Data Analysis

- Found DSSTox matches for 7011 substances
- Extracted MW values



GenRA approach : Overall 'global' performance

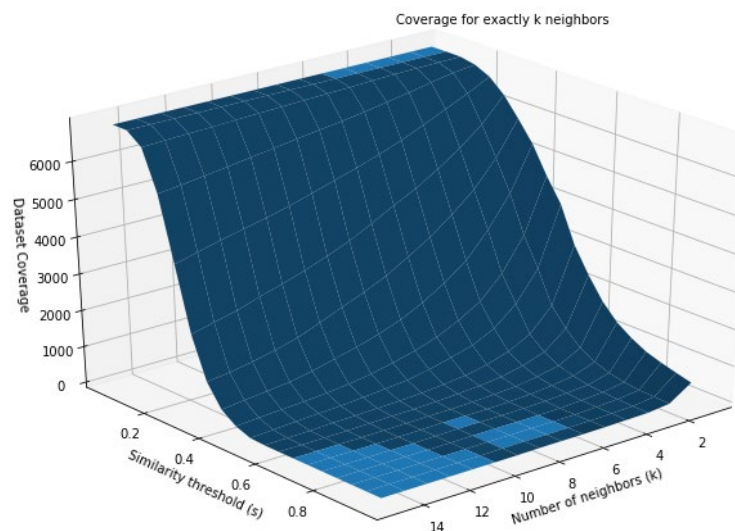
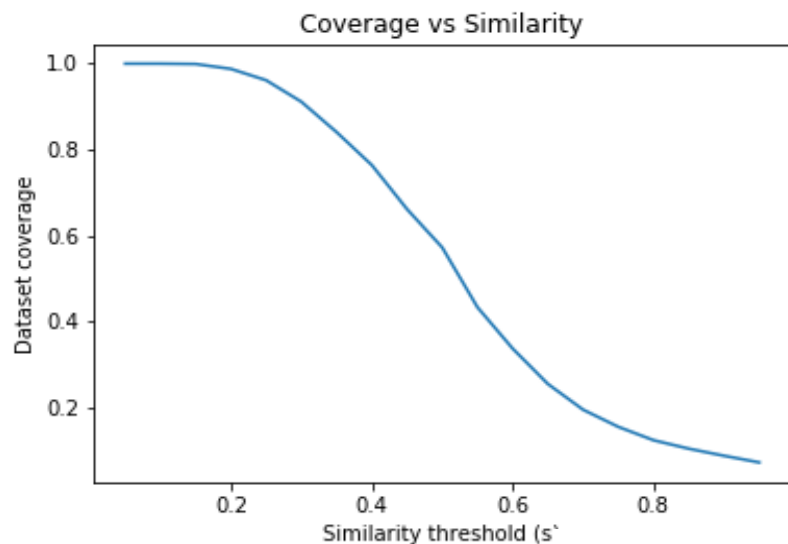
- Search for a maximum of 10 nearest neighbours on entire dataset
- Use a min similarity threshold of 0.5



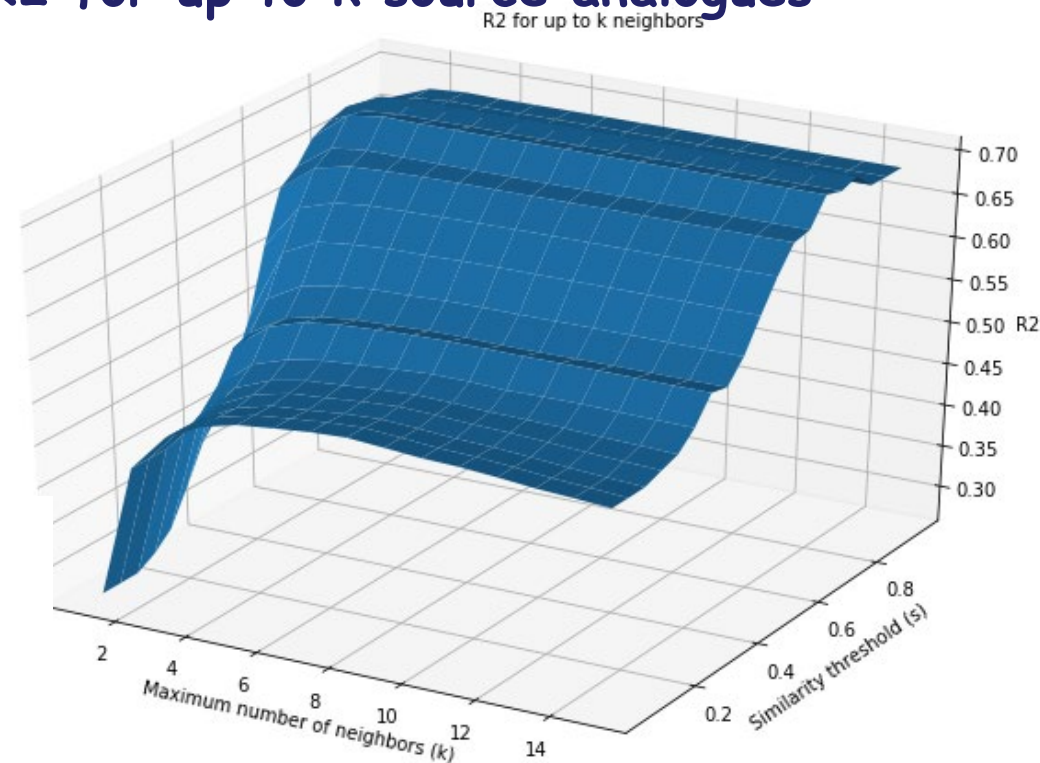
- Linear regression used to fit predicted and observed LD50 values
- $R^2 = 0.61$
- RMSE = 0.58
- A few outliers, but not too extreme
- Residuals clustered around zero with no obvious patterns

Coverage vs Similarity vs Performance

- Coverage vs Similarity**

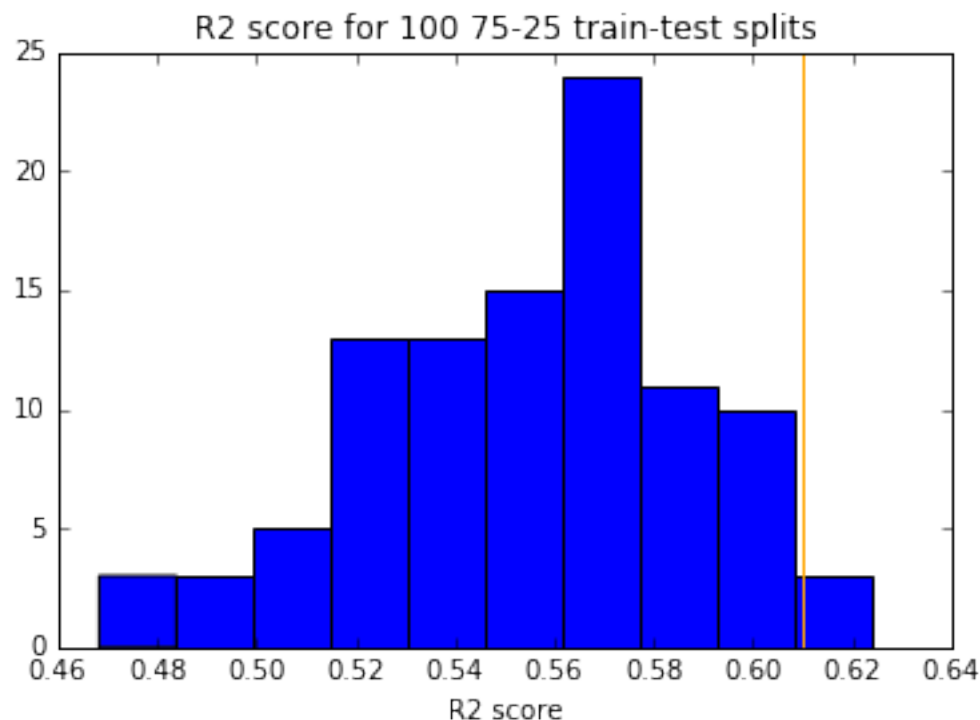


R² for up to k source analogues



Based on the grid searches performed, $k = 10$, $s = 0.5$ were reasonable parameters to tradeoff coverage vs prediction accuracy

Monte Carlo Cross Validation



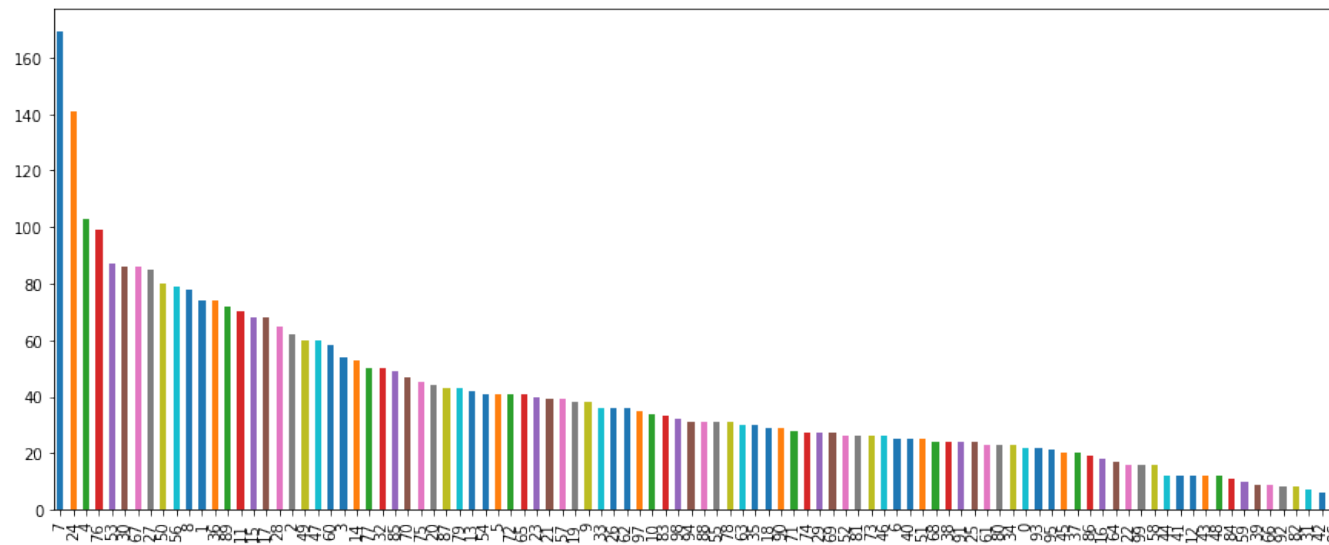
- Estimate confidence in R2
- 75-25 train-test splits
- R^2 values range from 0.46 to 0.62
- *GenRA* performs strongly and robustly on this acute tox data set.

Evaluating 'local' performance

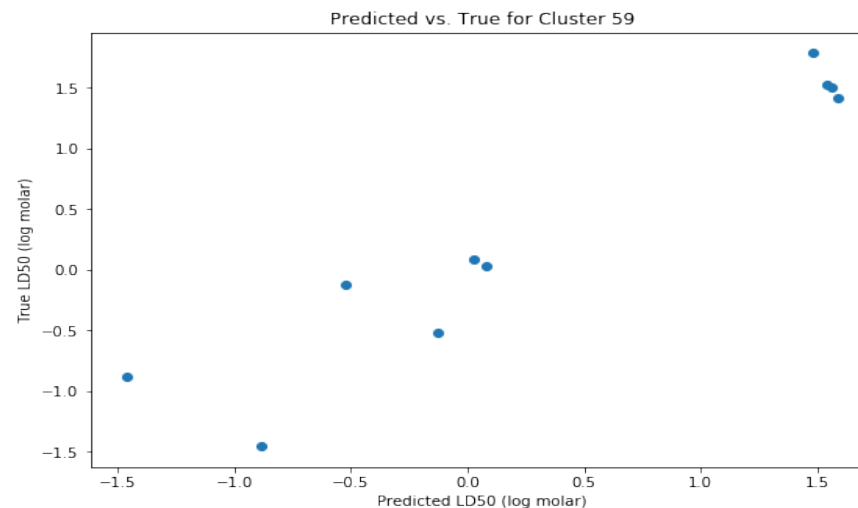
Clustered chemicals into 100 groups on the basis of ToxPrint fingerprints

Explored performance on the basis of individual clusters to gauge what sorts of chemicals resulted in significantly improved performance (R^2) relative to the overall 'global' performance reported using 10 nearest neighbours and a similarity of 0.5

Average R^2 values improved ($R^2 > 0.61$) for 19 out of the 100 clusters, some up to 0.91

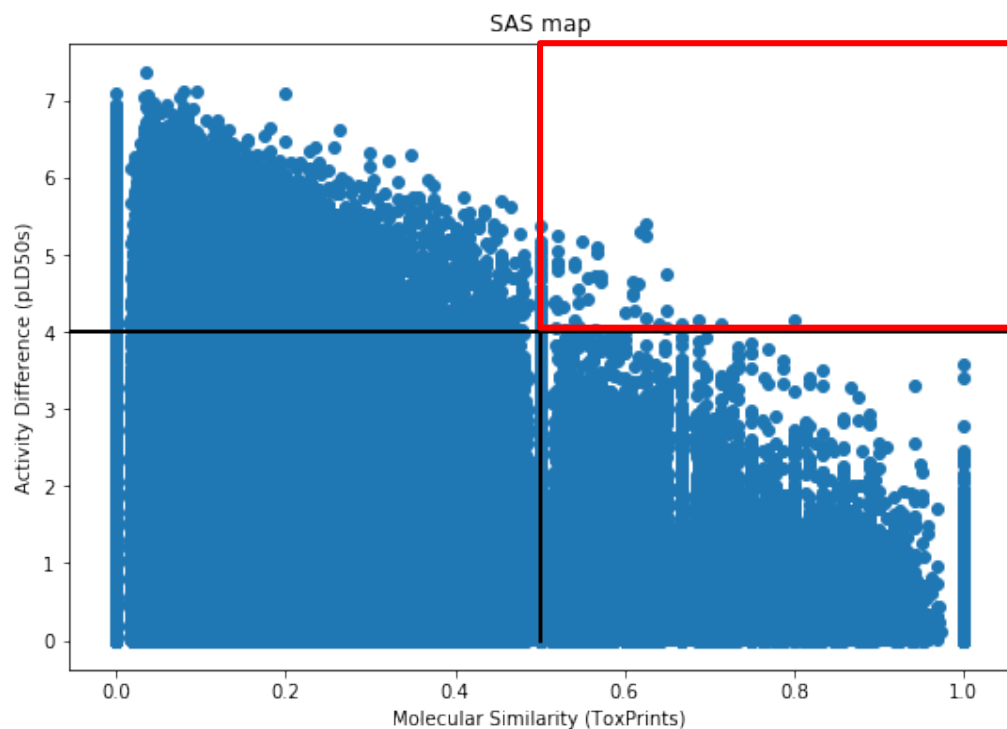


Carbamate containing substances



Structure-Activity similarity (SAS) map

- Are there pairs of substances that are very similar structurally with very high LD50 differences, so called activity cliffs



The number of chemical pairs that fell within the activity cliff quadrant was very low relative to the total number of chemical pairs captured.



This suggests that the chemical fingerprints were able to capture sufficient information to make robust predictions of acute oral toxicity.

Take home messages

- Initial GenRA (baseline) considered structural similarity and/or bioactivity to make binary predictions of toxicity
- Recent work has transitioned towards extending the GenRA approach to make quantitative predictions of toxicity
- This case study used the acute oral toxicity LD50 values collected as part of the ICCVAM ATWG and applied it to GenRA
- Using chemical fingerprints alone, a reasonable fit of R^2 of 0.61 using k up to 10 and $\min s$ of 0.5
- This was a pragmatic set of parameters to balance performance with coverage
- On a 'local' level, 19 out of 100 clusters of chemicals were found to show much improved performance (up to a R^2 of 0.91 in certain cases)

Acknowledgements

- Many but in particular..
- George Helman
- Imran Shah
- Tony Williams
- Jeff Edwards
- Jason Lambert
- Lucy Lizarraga
- Agnes Karmaus
- Nicole Kleinstreuer