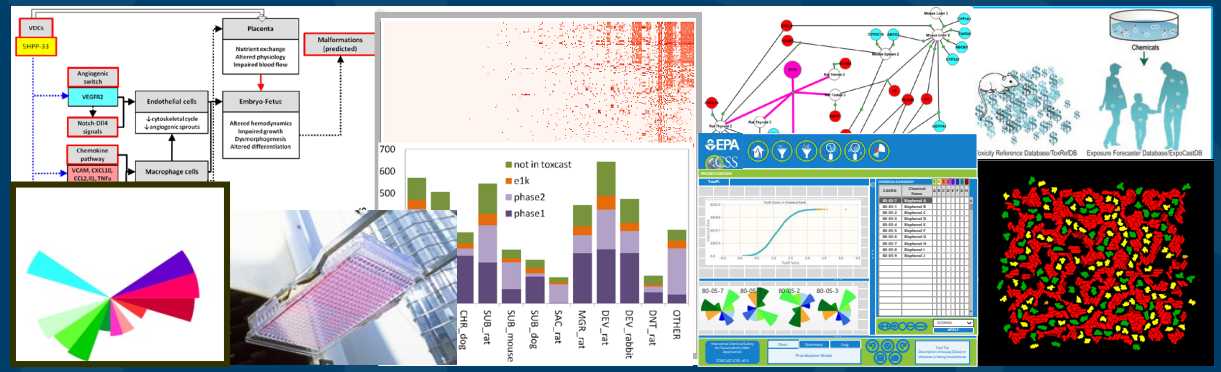
Transitioning GenRA to Quantitative Predictions: A Case Study of Acute Toxicity Data



George Helman[1,2], Imran Shah[2], Grace Patlewicz[2]

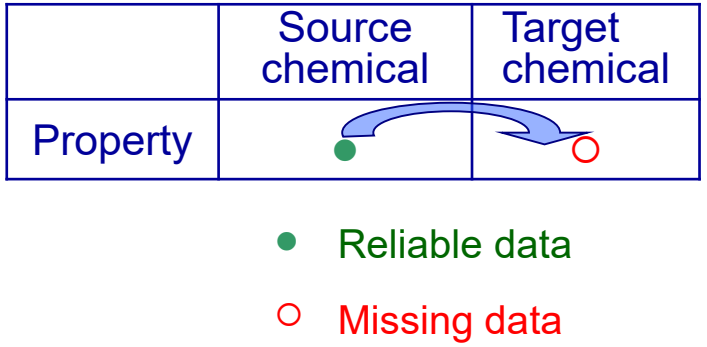
[1]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA
[2]National Center for Computational Toxicology, US EPA, RTP, NC, USA

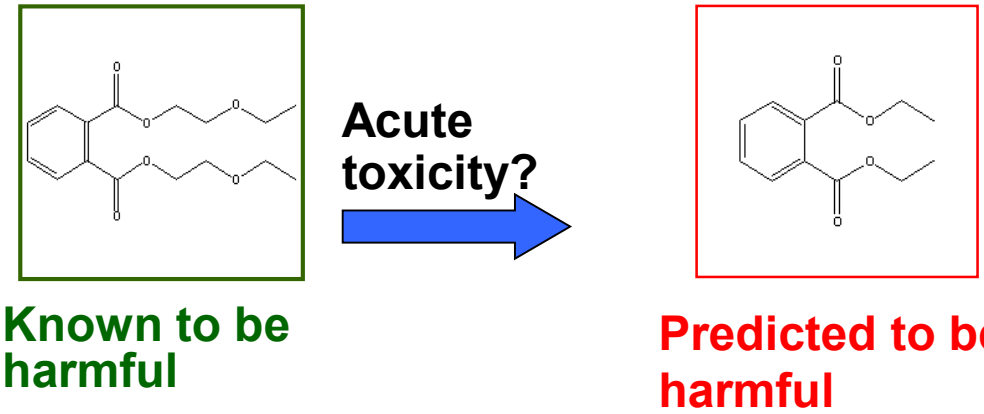
Outline

- Overview of GenRA
- Using GenRA for acute toxicity point of departure prediction
- Evaluation of predictions
- Future work + conclusions

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.





GenRA - Introduction

- GenRA (Generalized Read-Across) is a “local validity” approach predicting toxicity as a similarity-weighted activity of source analogues based on chemistry and/or bioactivity descriptors. (Shah et al, 2016)

- Generalized version of 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.

Methods

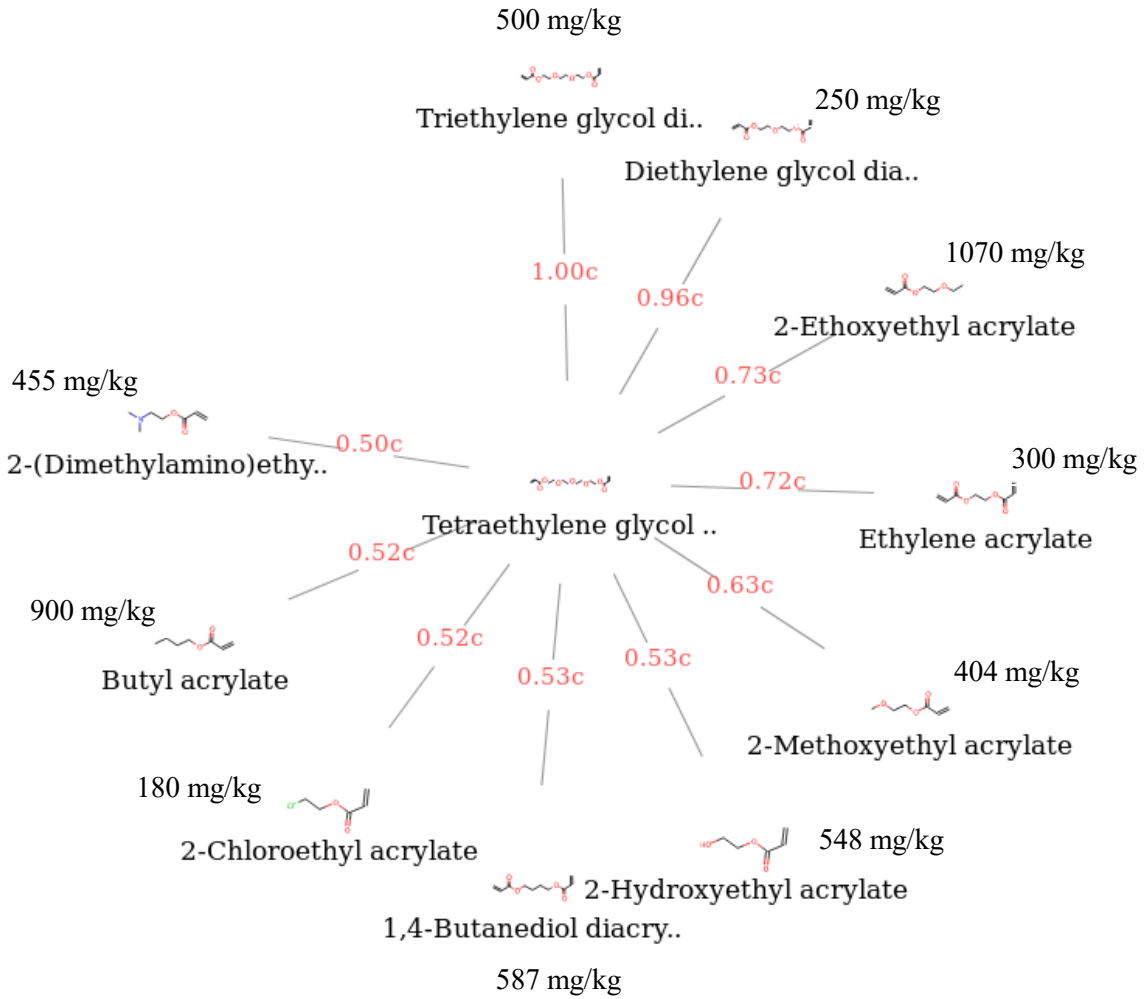
- GenRA is a similarity-weighted activity score of nearest neighbors

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

- Similarity calculated using Jaccard distance over Morgan chemical fingerprints
- Search for a maximum of 10 nearest neighbors on entire dataset.
- Use a similarity threshold of 0.5

Example

- Target: Tetraethylene glycol diacrylate
- Molecular weight: 235.06
- Calculate similarity between tetraethylene glycol diacrylate and every other chemical in the dataset based on Jaccard distance of Morgan chemical fingerprints.
- Select 10 chemicals with highest similarities (shown to right)



Calculating Example Prediction

Name	Similarity	LD50 value (mg/mol)
Tetraethylene gylcol diacrylate	1.0 (target)	-0.430
Triethylene glycol diacrylate	1	-0.287
Diethylene glycol diacrylate	0.96	-0.067
2-Ethoxyethyl acrylate	0.73	-0.871
Ethylene acrylate	0.72	-0.246
2-Methoxyethyl acrylate	0.63	-0.492

- $y_{\text{tetraethylene glycol diacrylate}}$ = 1*-0.287 + 0.96*-0.067 + 0.73*-0.871 + 0.72 * -0.246 + 0.63 * -0.492 / 1 + 0.96 + 0.73 + 0.72 + 0.63
  = -0.428 (log molar)
- mg/kg prediction: 10^(-(-0.428))*235.06 = 809.539 mg/kg

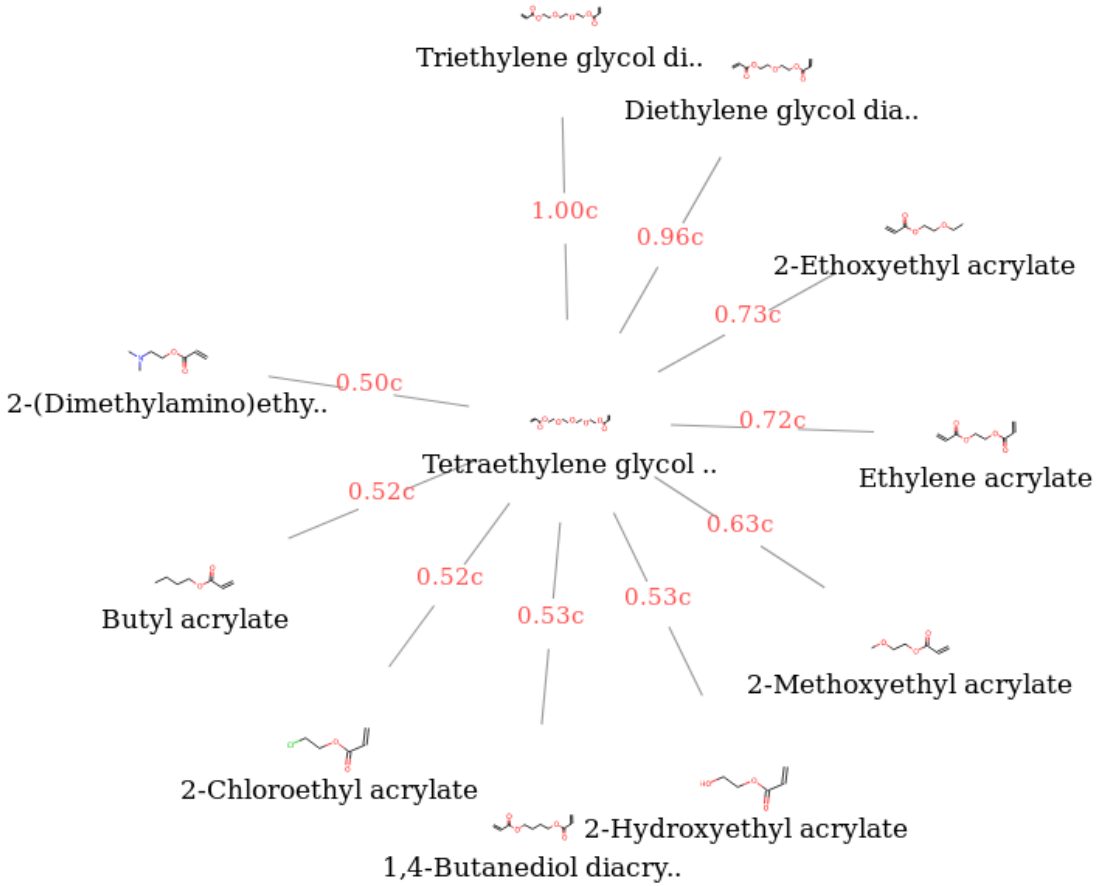
Example Predictions

Log Molar

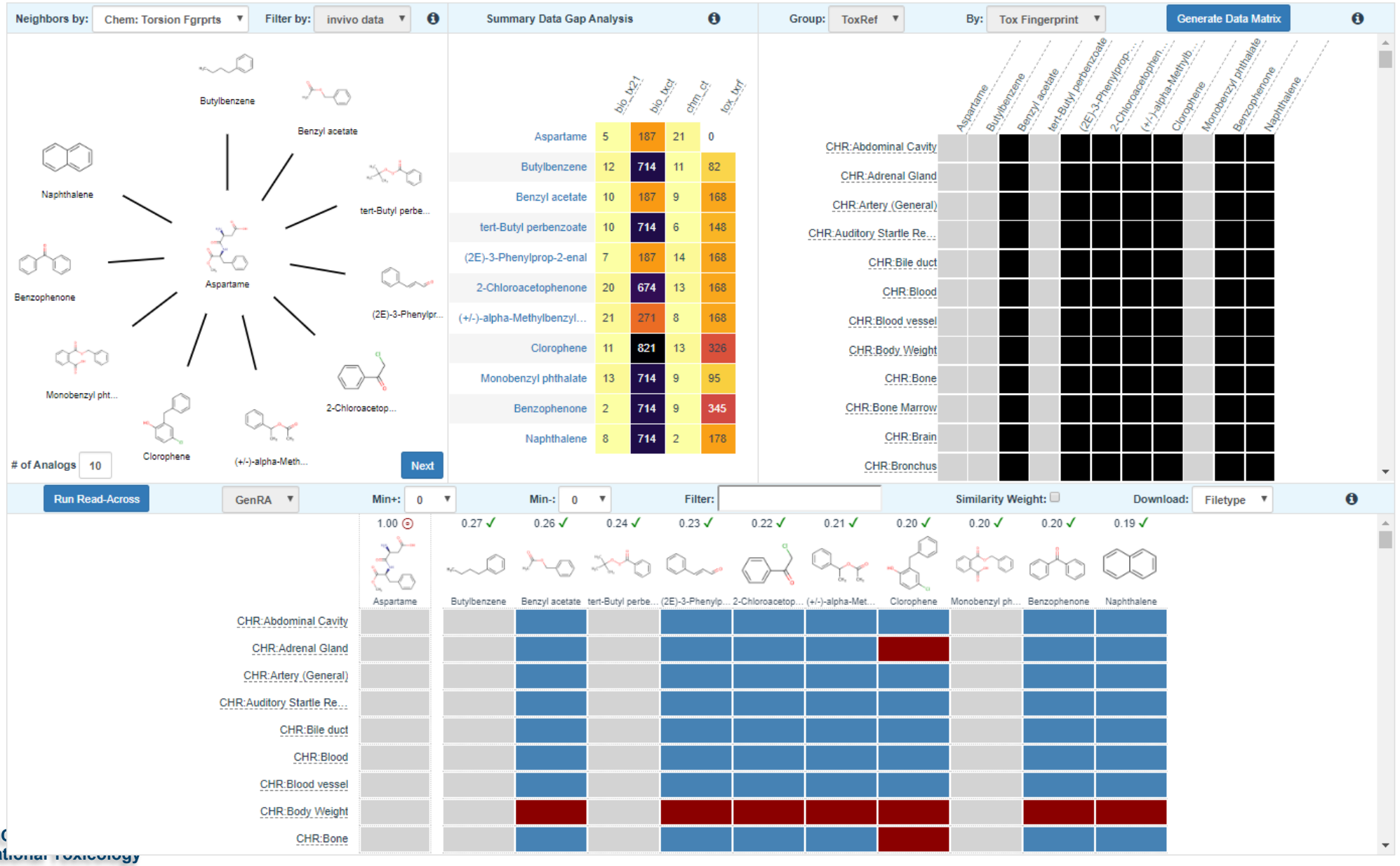
- Predicted value: -0.428
- Actual value: -0.430
- Residual: 0.002

Mg/kg

- Predicted value: 809.539
- Actual value: 813
- Residual: 3.461



Web-based Workflow



Original Application

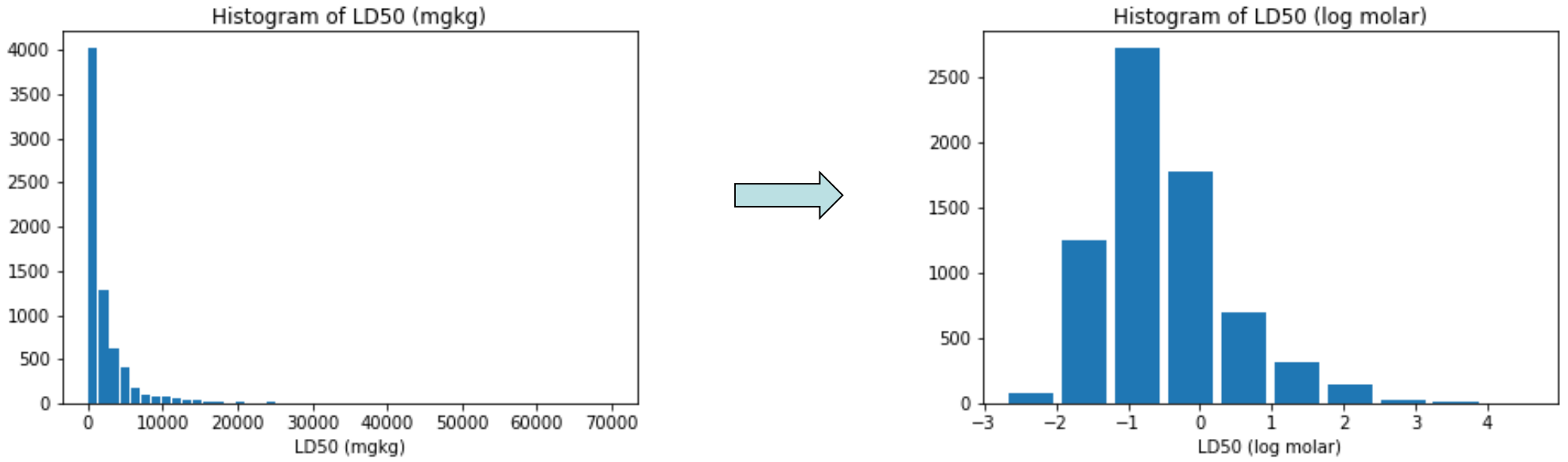
- Underlying data used was taken from ToxRefDB, a collection of repeated dose toxicity study types e.g. chronic, multigeneration, developmental, subchronic etc
- Toxicity effects within those study types were recorded as binary outcomes (0 for non-toxic, 1 for toxic)
- Toxicity effects were then predicted as binary outcomes (0 or 1)
- Dataset was clustered into local validity domains to find areas of chemical space where method performs best

EPA United States Environmental Protection Agency

Current Application

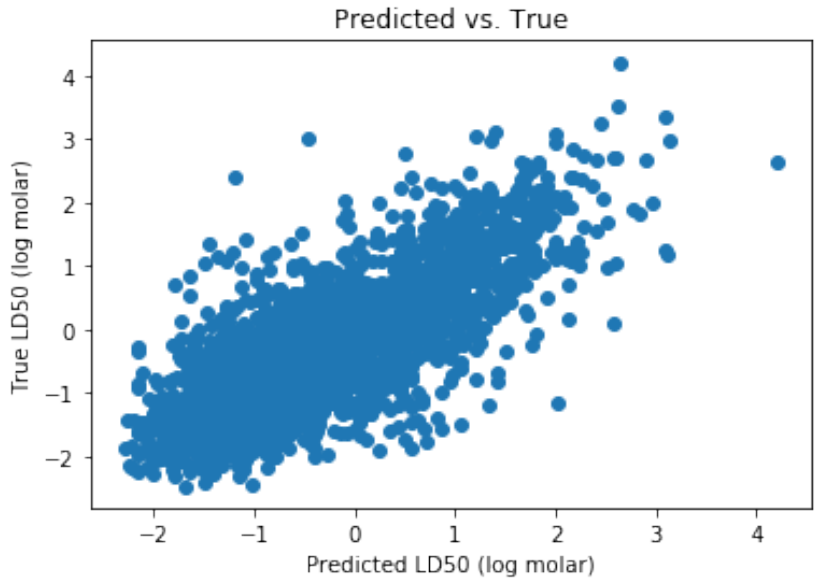
- We would like to test how GenRA performs on non-binary data.
- Acute rat oral toxicity (LD50) dataset with 16173 assays of 11992 substances
- Found DSSTox matches for 9293 substances (13295 assays)
- Median of the lowest quartile after removal of extreme values used for substances with multiple studies

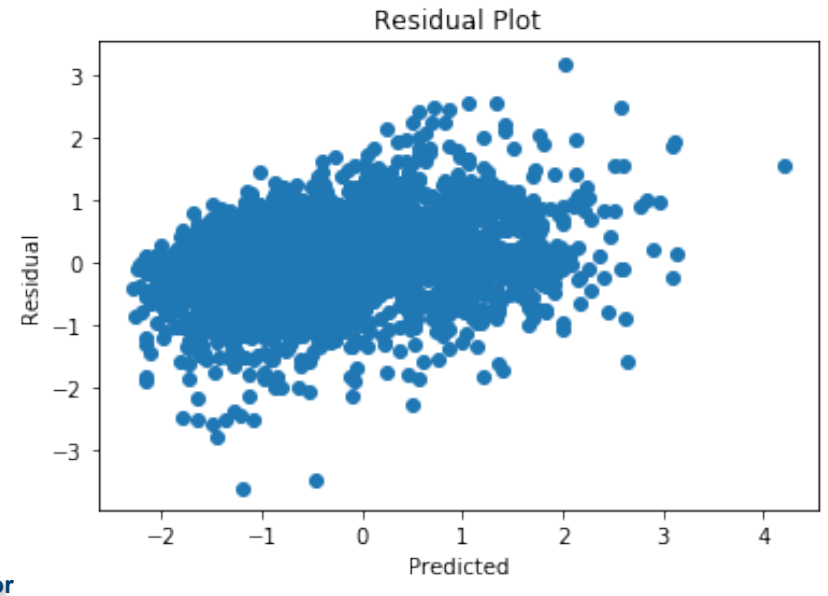
Exploratory Data Analysis



- Untransformed data highly skewed with extreme outliers
- Log molar transformation looks approximately normal

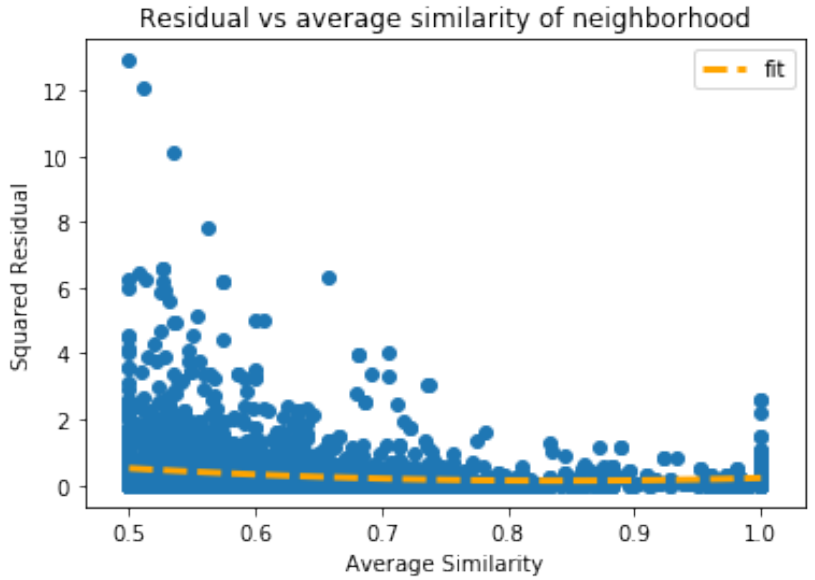
Fit

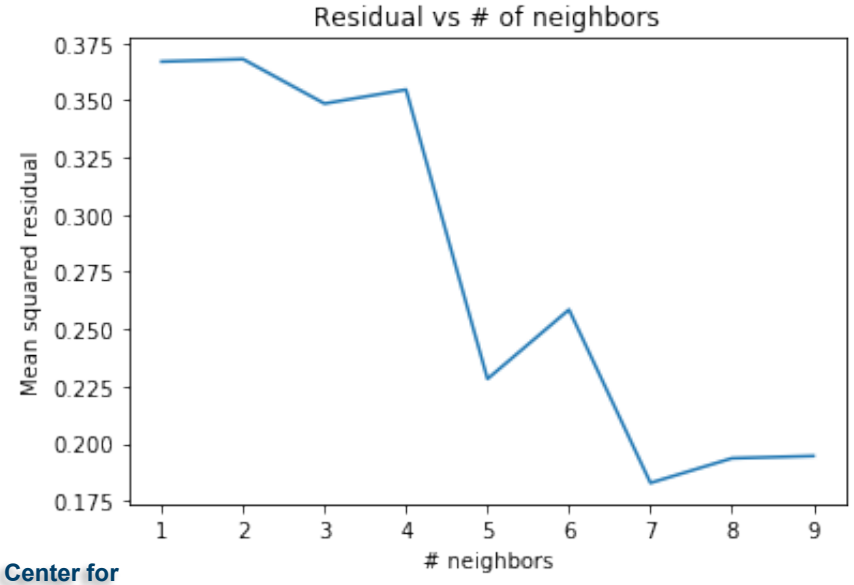




- R^2 = 0.61
- RMSE = 0.58
- A few outliers, but not too extreme
- Residuals clustered around zero with no obvious patterns

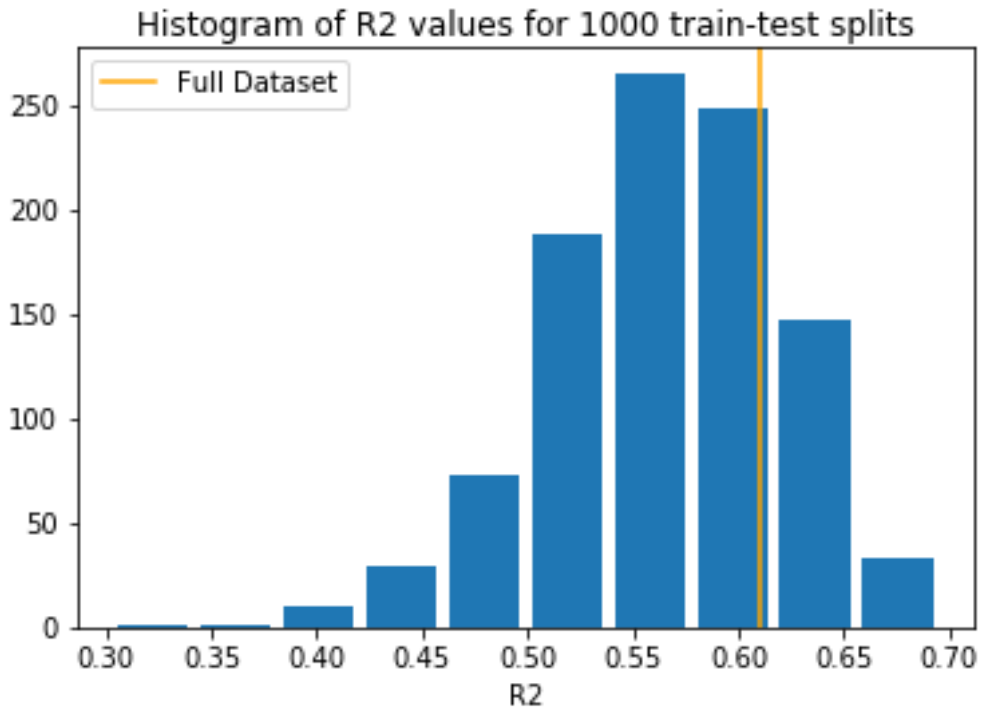
Fit, cont.





- Outliers tend to be for dissimilar neighborhoods
- Increasing similarity of the neighborhood leads to better predictions
- More neighbors in the neighborhood also leads to better predictions.

Evaluation of GenRA Performance



- 90-10 train-test splits
- R^2 values range from 0.30 to 0.70
- We select up to 10 nearest neighbors for each target chemical
- s=0.5 is threshold of similarity under which

Comparison to Other methods

- Zhu et al 2009
  - Rat LD50 data for 361 chemicals
  - k-Nearest Neighbors after partitioning data into 2 classes based on lc50 vs ld50 relationship
  - $0.50 < R^2 < 0.60$ (cross-validation)
- Alberga et al 2018:
  - Rat LD50 data for 8944 chemicals
  - k-Nearest Neighbors using combination of 19 fingerprints
  - $R^2 = 0.723$ (best)
- GenRA
  - Rat LD50 data for 7011 chemicals
  - k-Nearest Neighbors with Morgan fingerprints
  - R2 = 0.61 (k=10 and s=0.5)

Future Work + Conclusions

- GenRA has previously been used to predict binary toxicity calls, but is now shown to be applicable to non-binary datasets.

- GenRA predicts LD50 values accurately on this dataset and these predictions are robust because it predicts well in cross-validation.

- Future work
  - Incorporate phys-chem to see if it improves performance, particularly for outliers
  - Perform analysis cluster-by-cluster using clusters identified in Shah et al, 2016
  - Add quantitative predictions to GenRA web tool