Published in final edited form as:
*Bioinformatics*. 2021 October 11; 37(19): 3380–3381. doi:10.1093/bioinformatics/btab210.

# Generalized Read-Across prediction using genra-py

**Imran Shah**[*],
**Tia Tate**,
**Grace Patlewicz**
Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, NC 27709, USA

## Abstract

**Motivation:** Generalized Read-Across (GenRA) is a data-driven approach to estimate physico-chemical, biological or eco-toxicological properties of chemicals by inference from analogues. GenRA attempts to mimic a human expert's manual read-across reasoning for filling data gaps about new chemicals from known chemicals with an interpretable and automated approach based on nearest-neighbors. A key objective of GenRA is to systematically explore different choices of input data selection and neighborhood definition to objectively evaluate predictive performance of automated read-across estimates of chemical properties.

**Results:** We have implemented genra-py as a python package that can be freely used for chemical safety analysis and risk assessment applications. Automated read-across prediction in genra-py conforms to the scikit-learn machine learning library's estimator design pattern, making it easy to use and integrate in computational pipelines. We demonstrate the data-driven application of genra-py to address two key human health risk assessment problems namely: hazard identification and point of departure estimation.

**Availability and implementation:** The package is available from github.com/i-shah/genra-py.

## 1 Introduction

Chemicals are evaluated for human and ecological health risk before commercial use (USEPA, 2018b, 1996). In the absence of experimental testing data for a new substance, 'read-across' techniques are employed to fill data by inference from a 'similar' substance or substances (OECD, 2017). The manual and potentially subjective nature of read-across makes it difficult to consistently perform, define the level of evidence required, and evaluate confidence or uncertainty in inferences made (Patlewicz *et al*., 2017). We developed Generalized Read-Across (GenRA) (Shah *et al*., 2016) for automated read-across prediction of a property of a chemical (called the target) as a similarity weighted average of the same property from analogues (also referred to as source analogues) (Low *et al*., 2013). To address the limitations of available read-across tools (Patlewicz *et al*., 2017), we developed a server-based version of GenRA to support interactive chemical safety analysis workflows

[*]To whom correspondence should be addressed: shah.imran@epa.gov.
*Conflict of Interest*: none declared.

on the web (Helman *et al.*, 2019a). Here we introduce genra-py, which is a stand-alone Python 3 package that implements GenRA functionality for integration into computational pipelines. With this package, it is possible to efficiently analyze thousands of chemicals in batch-mode or individual chemicals interactively via Jupyter notebooks.

## 2 Software

GenRA estimates a binary or continuous property ($\beta$) for a given chemical ($c_i$) based on the similarity ($S$) weighted activity of k-nearest neighbors ($\{c_{i1}, c_{i2}, . . . , c_{ik}\}$) as: $y^{\beta} = \frac{\sum s_{ij} x_i^{\beta}}{\sum s_{ij}}$ (Shah *et al.*, 2016). We assume $\boldsymbol{x}^{\beta}$ is a vector of property values for neighboring chemicals and $S_{ij}$ is the similarity between chemicals $c_i$ and $c_j$ calculated on the basis of input vectors $x_i^{\alpha}$ and $x_j^{\alpha}$, respectively. The input vectors $\boldsymbol{x}^{\alpha}$ are defined by chemical structure descriptors or other experimentally measured bioactivity data (where descriptor type, $\alpha \in \{$*chemical, bioactivity, etc.*$\}$). We implemented read-across inference of $y^{\beta}$ using $k$, $\boldsymbol{x}^{\alpha}$ , $S$ and $\boldsymbol{x}^{\beta}$ as two types of estimators: a classifier for categorical data (e.g. molecular bioactivity or hazard classes) and a regressor for continuous data (e.g. *in vitro* bioactivity concentrations, oral doses, etc.). The classifier and regressor were implemented in Python 3 using the k-nearest neighbors estimators in the scikit-learn library (Pedregosa *et al.*, 2011) as GenRAPredClass and GenRAPredValue, respectively. We implemented chemical structural visualizations using the RDKit cheminformatics library (Landrum, 2015). The source code and installation instructions are available from github.com/i-shah/genra-py.

## 3 Human health risk assessment use-cases

We present two use cases of genra-py to illustrate a data-driven approach to aiding two key steps in human health risk assessment, including hazard identification and point of departure (POD) estimation. Hazard identification determines whether exposure to a chemical causes any adverse health effects (based on animal testing data). In POD estimation, the severity of the effect (response) is related to the exposure (dose) using a quantitative dose-response assessment. Machine learning approaches use non-animal data (e.g. chemical structure, in vitro bioactivity, etc.) to frame hazard identification and POD estimation as classification and regression problems, respectively. The GenRA workflow for hazard identification and POD estimation involves the following main steps: (i) retrieving and loading data ($\boldsymbol{x}^{\alpha}$ and $\boldsymbol{x}^{\beta}$) for chemicals; (ii) selecting a property $b$ for prediction; (iii) evaluating predictive performance using cross-validation testing; and (iv) visualizing the structures for the $k$ source analogues of a target chemical as well as predictions $y^{\beta}$. The complete workflows and detailed documentation are provided as Jupyter notebooks and are available from github.com/i-shah/genra-py under the notebooks directory.

### 3.1 Hazard identification: repeat-dose testing

We use GenRAPredClass for hazard identification using a dataset consisting of 600 chemicals and 564 types of effects ($\beta$) observed in repeat-dose toxicity testing studies (Shah *et al.*, 2016). We systematically calculated F1 scores using 5-fold cross-validation testing for each $\beta$ using different descriptor types ($\alpha \in \{$chemical, bioactivity, hybrid$\}$). We also conducted grid search for the optimal number of $k \in \{1, 2, 3, \ldots, 20\}$. Lastly, we illustrated the application of GenRAPredClass to estimate the hazard classes for an untested chemical (nitrofen). The complete analysis is available in the Jupyter notebook: 010-genra-py-shah-2016.ipynb. The entire analysis took 46 s to execute on 16 × 86 cores at 3.7 GHz.

### 3.2 POD estimation: acute toxicity

Next, we showcased GenRAPredValue for estimating rat 50% lethal dose (LD50) in acute toxicity studies using published data (Helman *et al.*, 2019b). First, we loaded the data and generated the Morgan chemical fingerprints using the RDKit library (Landrum, 2015). Second, we systematically evaluated coefficient of determination (R2) scores using 5-fold cross-validation testing to estimate LD50 values. Lastly, we calculated the LD50 value for each chemical using all other chemicals and compared them with the actual values. The complete analysis is available in the Jupyter notebook: 010-genrapy-helman-2019.ipynb. This analysis took 60 s to execute on 16 × 86 cores at 3.7 GHz.

## 4 Summary

There is an urgent need for new approach methodologies to aid human health risk assessment (USEPA, 2018a). Only a small fraction of the chemicals in commerce have been thoroughly evaluated in traditional toxicity testing studies. GenRA embodies read-across best-practices that have emerged after much international deliberation and consensus in the risk assessment community (Patlewicz *et al.*, 2018; Rovida, 2020). It employs nearest neighbor inference, which is readily interpretable and mimics traditional read-across techniques. We hope the public release of the genra-py package and the use-cases described herein will stimulate new research on chemical risk assessment that is data-driven, automated, objective, transparent and reproducible.

## Acknowledgements

The authors acknowledge contributions from George Helman, and Russell Thomas to the GenRA approach over the years and Jason Lambert and Lucina Lizarraga for providing valuable feedback.

### Funding

This work was supported by the Office of Research and Development in the United States Environmental Protection Agency.

## References

Helman G et al. (2019a) Generalized Read-Across (GenRA): a workflow implemented into the EPA CompTox Chemicals Dashboard. ALTEX, 36, 1–5.

Helman G et al. (2019b) Transitioning the generalised read-across approach (GenRA) to quantitative predictions: a case study using acute oral toxicity data. Comput. Toxicol, 12, 100097.

Landrum G (2015) RDKit: Open-source cheminformatics; http://www.rdkit.org.

Low Y et al. (2013) Integrative chemical–biological read-across approach for chemical hazard classification. Chem. Res. Toxicol, 26, 1199–1208. [PubMed: 23848138]

OECD. (2017) Guidance on Grouping of Chemicals, Second Edition OECD.

Patlewicz G et al. (2018) Navigating through the minefield of read-across frameworks: a commentary perspective. Comput. Toxicol, 6, 39–54.

Patlewicz G et al. (2017) Navigating through the minefield of read-across tools: a review of in silico tools for grouping. Comput. Toxicol, 3, 1–18. [PubMed: 30221211]

Pedregosa F et al. (2011) Scikit-learn: machine Learning in Python. J. Mach. Learn. Res, 12, 2825–2830.

Rovida C (2020) Internationalization of read-across as a validated new approach method (NAM) for regulatory toxicology. ALTEX, 37, 579–606. [PubMed: 32369604]

Shah I et al. (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. [PubMed: 27174420]

USEPA. (1996) Summary of the Federal Insecticide, Fungicide, and Rodenticide Act. United States Environmental Protection Agency, Washington, DC.

USEPA. (2018a) Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program. pp. 1–39. United States Environmental Protection Agency, Washington, DC.

USEPA. (2018b) The Frank R. Lautenberg chemical safety for the 21st century act. pp. 1–67. United States Environmental Protection Agency, Washington, DC.