The TAME Toolkit for Introductory Data Science, Chemical-Biological Analyses, Predictive Modeling, and Database Mining for Environmental Health Research

Kyle Roell, Lauren Koval, Rebecca Boyles, Grace Patlewicz, Caroline Ring, Cynthia Rider, Cavin Ward-Caviness, David M. Reif, Ilona Jaspers, Rebecca C. Fry, and Julia E. Rager

Fall 2021

### Background Research in exposure science, toxicology, and environmental health is becoming increasingly reliant upon data science and computational methods that can more efficiently extract information from complex datasets. These methods can be leveraged to better identify relationships between exposures to chemicals in the environment and human disease outcomes. Still, there remains a critical gap surrounding the training of researchers on these in silico methods.

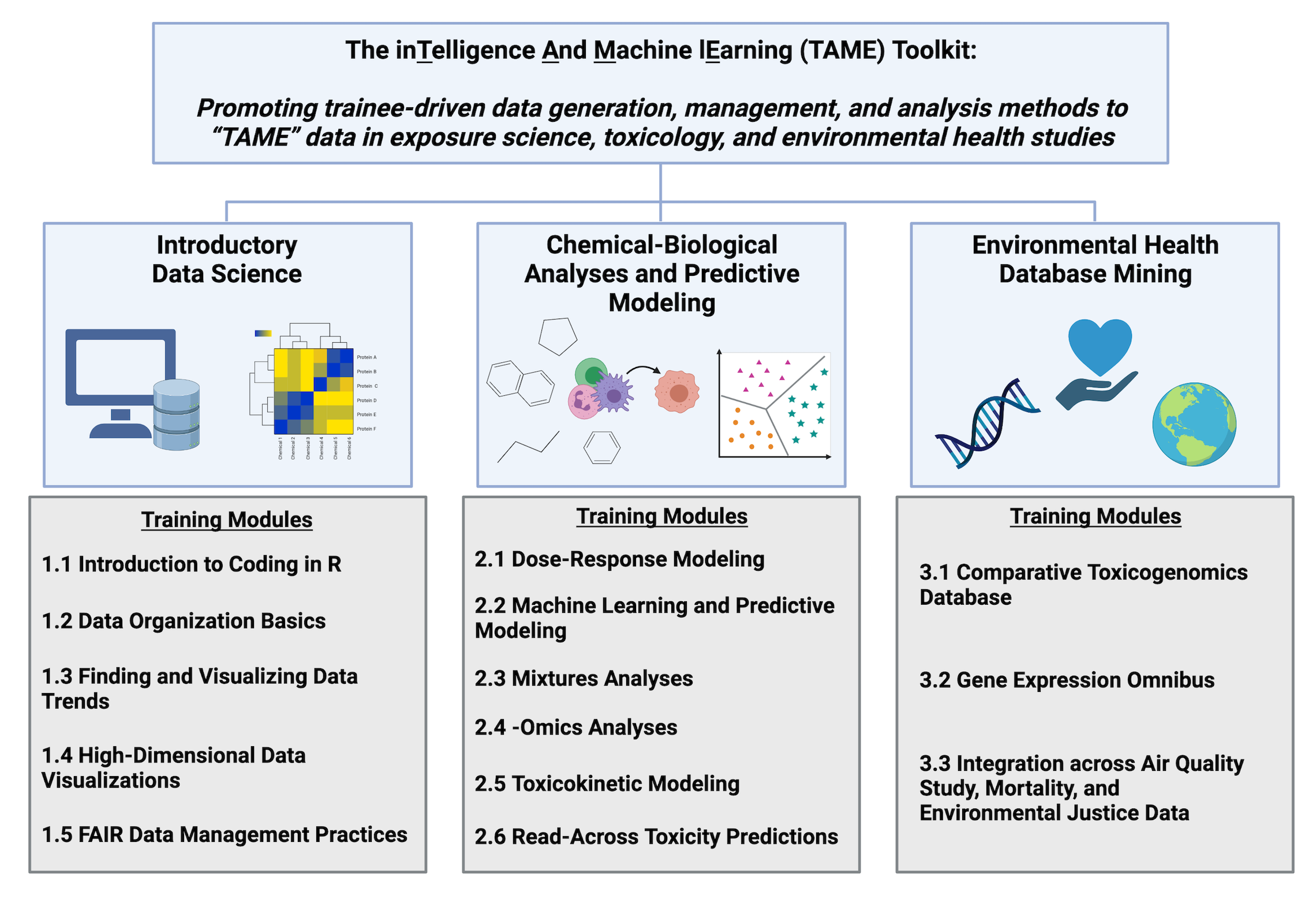
### Objectives

We aimed to address this critical gap by developing the in**T**elligence **A**nd **M**achine l**E**arning (TAME) Toolkit, promoting trainee-driven data generation, management, and analysis methods to “TAME” data in environmental health studies. This toolkit encompasses training modules, organized as chapters within this [Github Bookdown site](https://uncsrp.github.io/Data-Analysis-Training-Modules/).

### Module Development Overview

Training modules were developed to provide applications-driven examples of data organization and analysis methods that can be used to address environmental health questions. Target audiences for these modules include students and professionals in academia, government, and industry that are interested in expanding their skillset. Modules were developed by study coauthors using annotated script formatted for R/RStudio coding language and interface and were organized into three chapters. The first group of modules focused on introductory data science, which included the following topics: setting up R/RStudio and coding in the R environment; data organization basics; finding and visualizing data trends; high-dimensional data visualizations with heat maps; and Findability, Accessibility, Interoperability, and Reusability (FAIR) data management practices. The second chapter of modules incorporated chemical-biological analyses and predictive modeling, spanning the following methods: dose-response modeling; machine learning and predictive modeling; mixtures analyses; -omics analyses; toxicokinetic modeling; and read-across toxicity predictions. The last chapter of modules was organized to provide examples on environmental health database mining and integration, including chemical exposure, health outcome, and environmental justice data.

The overall organization of this TAME toolkit is summarized below. Modules are organized into three chapters, that are listed on the left side of this website.



All underlying code (in RMarkdown), input files, and imported graphics for these modules can be found at the parent [UNC-SRP Github Page](https://github.com/UNCSRP).

### Concluding Remarks

Together, this valuable resource provides unique opportunities to obtain introductory-level training on current data analysis methods applicable to 21st century exposure science, toxicology, and environmental health. These modules serve as applications-based examples on how to “TAME” data within the environmental health research field, expanding the toolbox for career development and cross-training of scientists in multiple specialties, as well as supporting the next generation of data scientists.

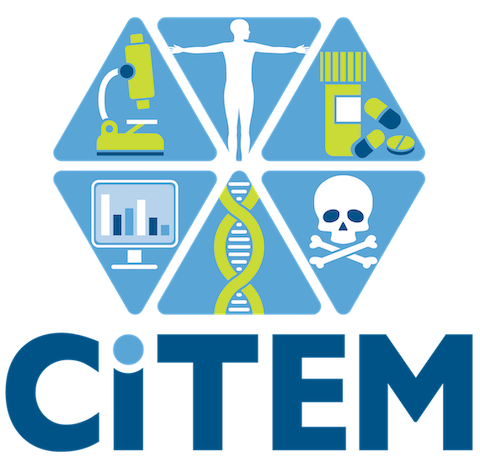
### Funding

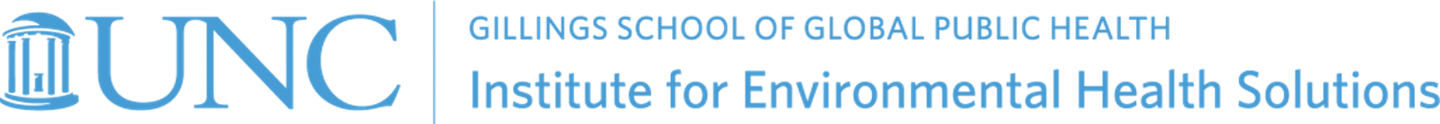
This study was supported by the National Institutes of Health (NIH) from the National Institute of Environmental Health Sciences, including the following grant funds and associated programs:

**P42ES031007**: The [University of North Carolina (UNC)-Superfund Research Program](https://sph.unc.edu/superfund-pages/srp/) (SRP) seeks to develop new solutions for reducing exposure to inorganic arsenic and prevent arsenic-induced diabetes through mechanistic and translational research. The UNC-SRP is Directed by Dr. Rebecca C. Fry.

The [UNC-SRP Data Analysis and Management Core (UNC-SRP-DMAC)](https://sph.unc.edu/superfund-pages/dmac/) provides the UNC-SRP with critical expertise in bioinformatics, statistics, data management, and data integration. Dr. Julia E. Rager is a Leader of the UNC-SRP-DMAC.

 **T32ES007126**: The [UNC Curriculum in Toxicology and Environmental Medicine (CiTEM)](https://www.med.unc.edu/toxicology/) seeks to provide a cutting edge research and mentoring environment to train students and postdoctoral fellows in environmental health and toxicology. Towards this goal, the CiTEM has a T32 Training Program for Pre- and Postdoctoral Training in Toxicology to support the development of future investigators in environmental health and toxicology. This training program has received supplement funds to expand training efforts centered on data management and data science practices to address current health issues in toxicology and environmental science. The UNC CiTEM is Directed by Dr. Ilona Jaspers.

 Support was additionally provided through the [Institute for Environmental Health Solutions (IEHS)](https://sph.unc.edu/iehs/institute-for-environmental-health-solutions/) at the University of North Carolina (UNC) Gillings School of Global Public Health. The IEHS is aimed at protecting those who are particularly vulnerable to diseases caused by environmental factors, putting solutions directly into the hands of individuals and communities of North Carolina and beyond. The IEHS is Directed by Dr. Rebecca C. Fry.



### Author Affiliations

Kyle Roell1,†, Lauren Koval1,2,†, Rebecca Boyles3, Grace Patlewicz4, Caroline Ring4, Cynthia Rider5, Cavin Ward-Caviness6, David M. Reif7, Ilona Jaspers1,2,8,9,10, Rebecca C. Fry1,2,8, and Julia E. Rager1,2,8,9

1The Institute for Environmental Health Solutions, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

2Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

3Research Computing, RTI International, Durham North Carolina, USA

4Center for Computational Toxicology and Exposure, US Environmental Protection Agency, Chapel Hill, North Carolina, USA

5Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

6Center for Public Health and Environmental Assessment, US Environmental Protection Agency, Chapel Hill, North Carolina, USA

7Department of Biological Sciences, and the Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA

8Curriculum in Toxicology and Environmental Medicine, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

9Center for Environmental Medicine, Asthma and Lung Biology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

10Department of Pediatrics, Microbiology and Immunology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

†These authors have contributed equally to this work and share first authorship