The inTelligence And Machine lEarning (TAME) Toolkit for Introductory Data Science in Environmental Health

Julia E. Rager, PhD, MSEE

Assistant Professor | University of North Carolina at Chapel Hill (UNC)

Department of Environmental Sciences and Engineering (ENVR)

Institute for Environmental Health Solutions (IEHS)

Center for Environmental Medicine and Lung Biology (CEMALB)

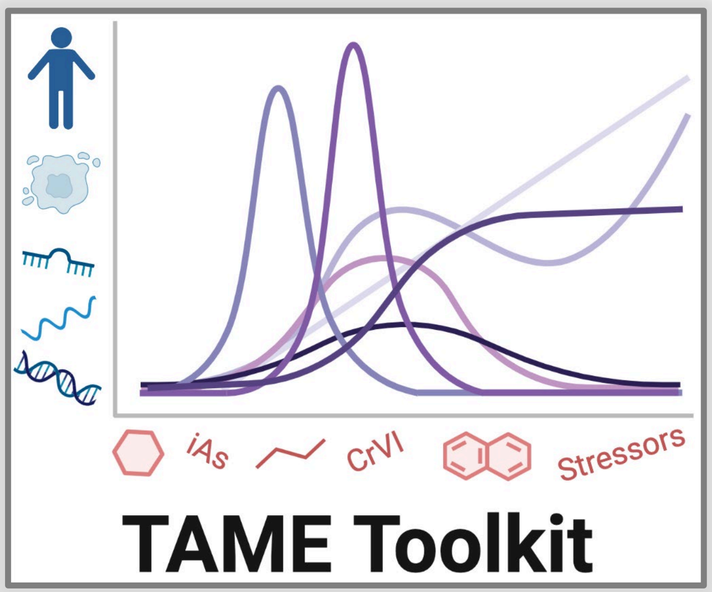
Curriculum in Toxicology and Environmental Medicine (CiTEM)

Kyle Roell, PhD

Data Analyst | University of North Carolina at Chapel Hill (UNC)

Department of Environmental Sciences and Engineering (ENVR)

Institute for Environmental Health Solutions (IEHS)

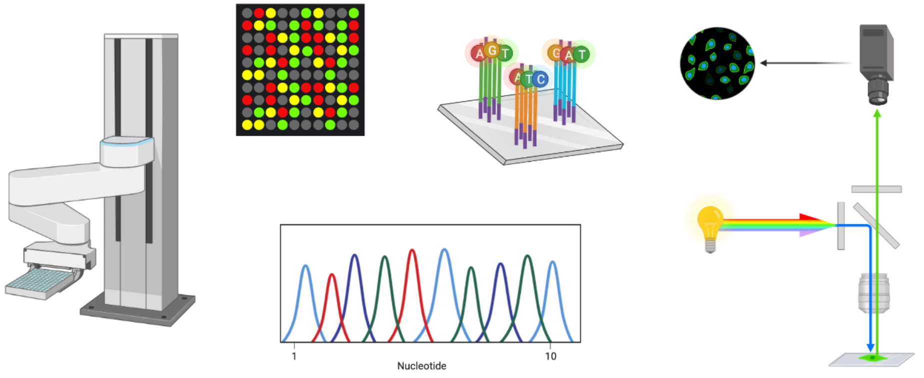


Presentation Schedule

- Introduction to TAME and its development – Dr. Julia Rager
- Real-time demonstration of how to use TAME – Dr. Kyle Roell
- Q&A is encouraged throughout and also at the end

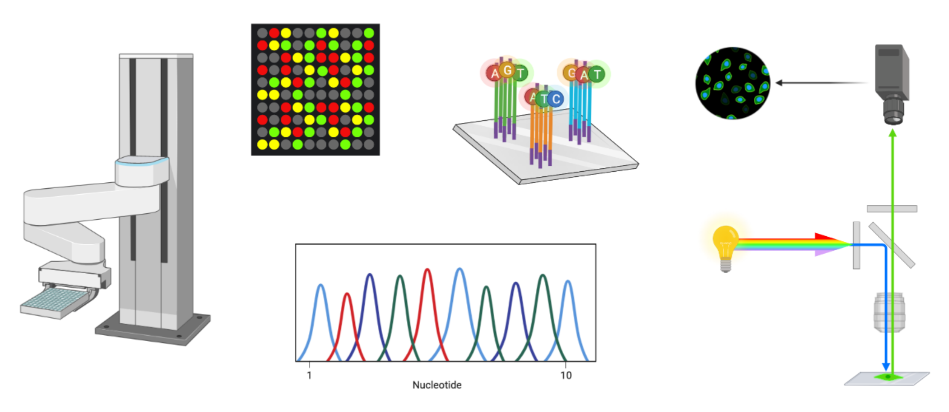
Need for Improved Data Science Training

High-Throughput / High-Content Technologies are Generating More Data



Need for Improved Data Science Training

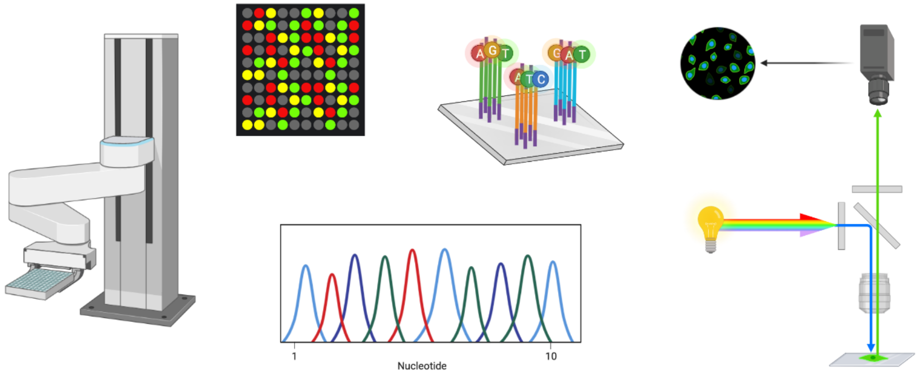
High-Throughput / High-Content Technologies are Generating More Data



Cross-Disciplinary Research is Becoming More Common, Integrating Multiple Variable Types

Need for Improved Data Science Training

High-Throughput / High-Content Technologies are Generating More Data



Cross-Disciplinary Research is Becoming More Common, Integrating Multiple Variable Types

Data Preservation Efforts and Record Keeping are Becoming Pillars to Scientifically Sound Research

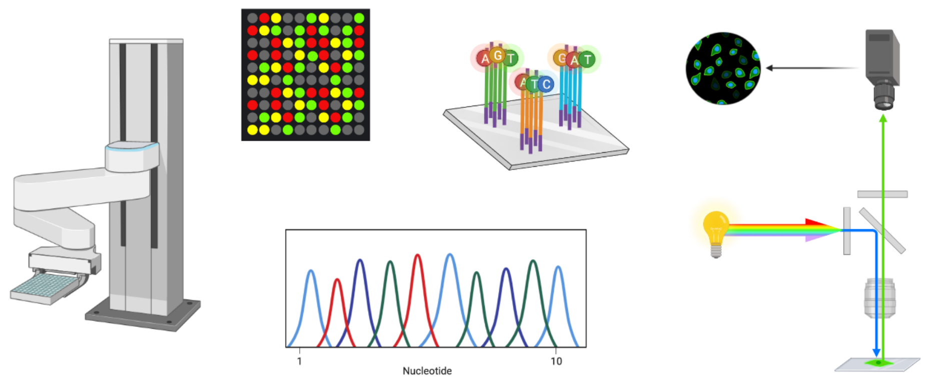
FAIR
(Findable, Accessible, Interoperable, Reusable) Guiding Principles for Scientific Data Management and Stewardship
(Wilkinson et al. 2016)

EPA's Scientific Data Management Plans (**SDMPs**)

NIH's Data Management Plans (**DMPs**) and Data Management Templates (**DMTs**)

Need for Improved Data Science Training

High-Throughput / High-Content Technologies are Generating More Data



Cross-Disciplinary Research is Becoming More Common, Integrating Multiple Variable Types

Data Preservation Efforts and Record Keeping are Becoming Pillars to Scientifically Sound Research

FAIR
(Findable, Accessible, Interoperable, Reusable) Guiding Principles for Scientific Data Management and Stewardship
(Wilkinson et al. 2016)

EPA's Scientific Data Management Plans (**SDMPs**)

NIH's Data Management Plans (**DMPs**) and Data Management Templates (**DMTs**)

How can we make use of these growing resources to better understand environmental stressors???

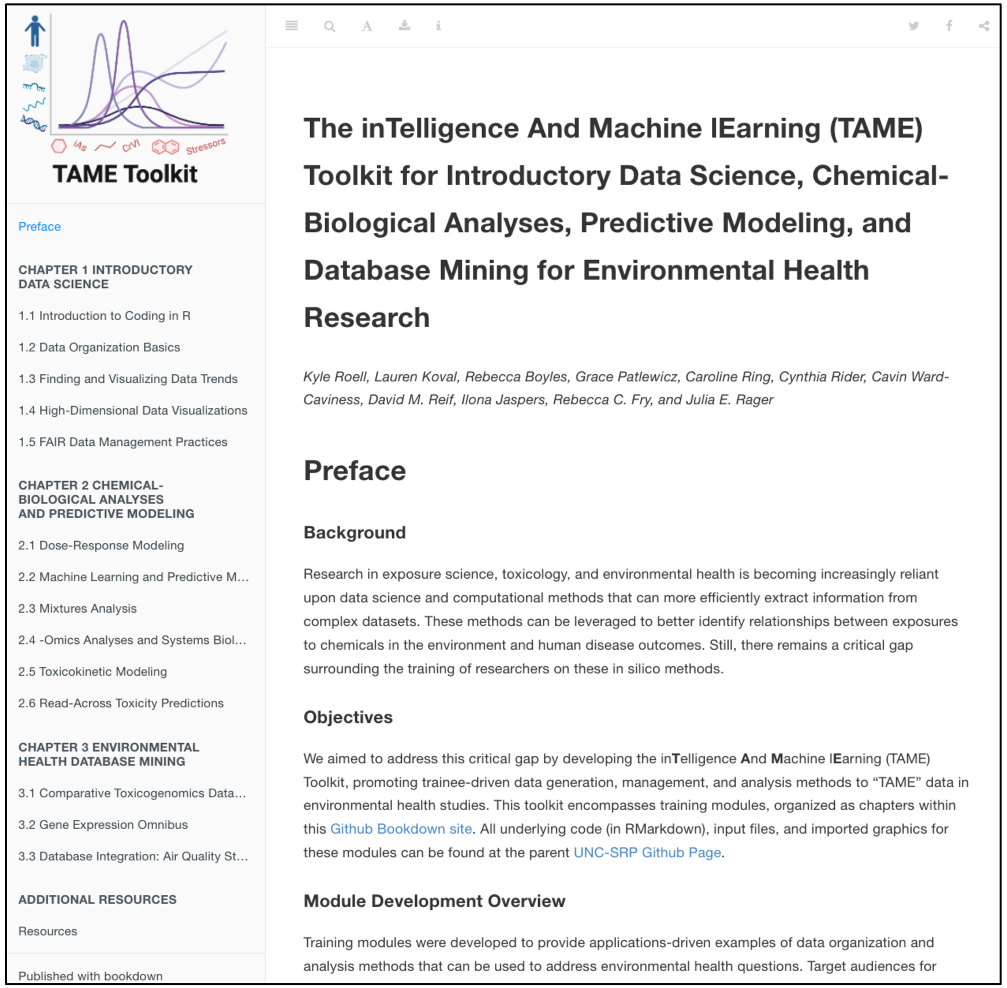
<u>DATA SCIENCE TRAINING</u>

Data Science Training through TAME Toolkit

- Multiple authors share knowledge and translate experiences gathered by multiple authors on data management and analysis methods to address real-world issues in environmental health through carefully developed training modules

- This training effort is currently being launched as the **inTelligence And Machine lEarning (TAME)** Toolkit

- Promotes didactic data generation, management, and analysis methods to **“TAME” data** in environmental health studies



https://uncsrp.github.io/Data-Analysis-Training-Modules/

https://github.com/UNCSRP

Wonderful Team of TAME Contributors

Kyle Roell
Data Analyst,
UNC-SRP, UNC

Lauren Koval
PhD Candidate,
Ragerlab, UNC

Rebecca Boyles
Data Management Director
and Scientist, RTI

Headshot not
preferred

Grace Patlewicz
Chemistry and
Cheminformatics Scientist,
US EPA

Caroline Ring
Computational Exposure
Scientist, US EPA

Cynthia Rider
Mixtures Toxicologist,
NIEHS / NTP

Cavin Ward-Caviness
Computational Biologist,
US EPA

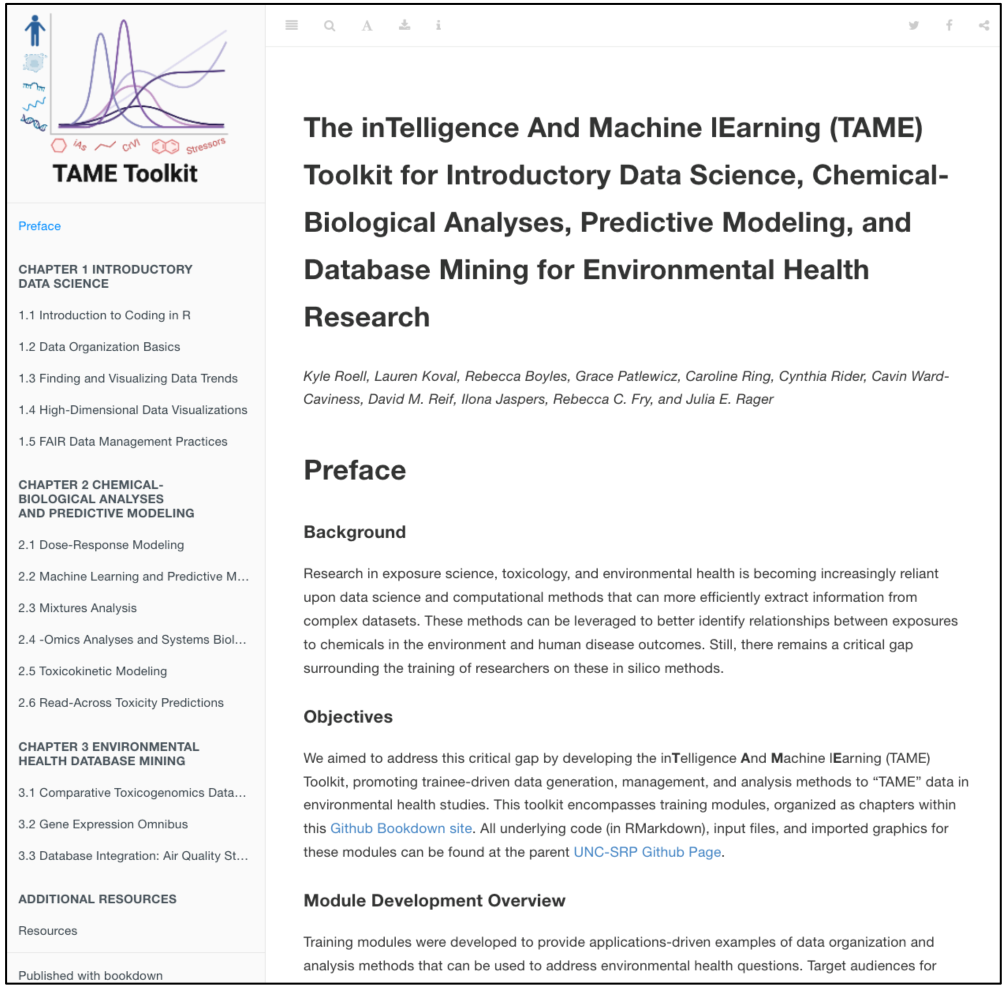
David Reif
Professor and NCSU-SRP
Data Management and Analysis
Lead, NCSU

Ilona Jaspers
Professor and Curriculum in
Environmental Medicine and
Toxicology Director, UNC

Rebecca Fry
Professor and Superfund
Research Program (UNC-SRP)
Director, UNC

Julia Rager
Assistant Professor and UNC-
SRP Data Management and
Analysis Co-Lead, UNC

TAME Modules Presented through Bookdown



TAME Toolkit

Published with bookdown

The inTelligence And Machine lEarning (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

Preface

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

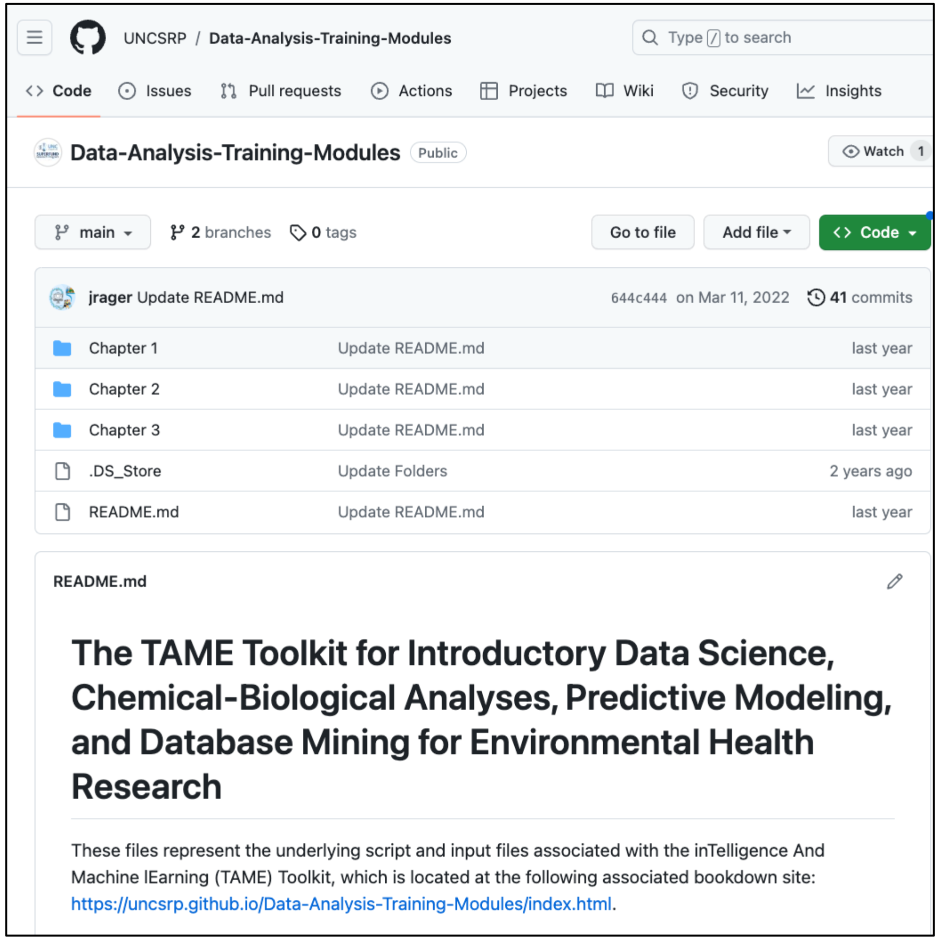
We aimed to address this critical gap by developing the in**T**elligence **A**nd **M**achine l**E**arning (TAME) Toolkit, promoting trainee-driven data generation, management, and analysis methods to "TAME" data in environmental health studies. This toolkit encompasses training modules, organized as chapters within this Github Bookdown site. All underlying code (in RMarkdown), input files, and imported graphics for these modules can be found at the parent UNC-SRP Github Page.

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

https://uncsrp.github.io/Data-Analysis-Training-Modules/

With underlying script and datasets posted on Github:



UNCSRP / **Data-Analysis-Training-Modules**

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights

Data-Analysis-Training-Modules Public

Watch 1

main | 2 branches | 0 tags | Go to file | Add file | Code

jrager Update README.md 644c444 on Mar 11, 2022 **41** commits

Chapter 1	Update README.md	last year
Chapter 2	Update README.md	last year
Chapter 3	Update README.md	last year
.DS_Store	Update Folders	2 years ago
README.md	Update README.md	last year

README.md

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

These files represent the underlying script and input files associated with the inTelligence And Machine lEarning (TAME) Toolkit, which is located at the following associated bookdown site: https://uncsrp.github.io/Data-Analysis-Training-Modules/index.html.

https://github.com/UNCSRP

Parent Manuscript

> Front Toxicol. 2022 Jun 22;4:893924. doi: 10.3389/ftox.2022.893924. eCollection 2022.

Development of the InTelligence And Machine LEarning (TAME) Toolkit for Introductory Data Science, Chemical-Biological Analyses, Predictive Modeling, and Database Mining for Environmental Health Research

Kyle Roell[1], Lauren E Koval[1,2], Rebecca Boyles[3], Grace Patlewicz[4], Caroline Ring[4], Cynthia V Rider[5], Cavin Ward-Caviness[6], David M Reif[7], Ilona Jaspers[1,2,8,9,10], Rebecca C Fry[1,2,8], Julia E Rager[1,2,8,9]

Affiliations + expand

PMID: 35812168 PMCID: PMC9257219 DOI: 10.3389/ftox.2022.893924

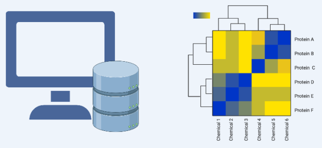
Free PMC article

Chapter—Based Content Organization

The inTelligence And Machine lEarning (TAME) Toolkit:

Promoting trainee-driven data generation, management, and analysis methods to "TAME" data in exposure science, toxicology, and environmental health studies

Introductory Data Science



Training Modules

1.1 Introduction to Coding in R

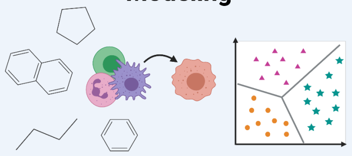
1.2 Data Organization Basics

1.3 Finding and Visualizing Data Trends

1.4 High-Dimensional Data Visualizations

1.5 FAIR Data Management Practices

Chemical-Biological Analyses and Predictive Modeling



Training Modules

2.1 Dose-Response Modeling

2.2 Machine Learning and Predictive Modeling

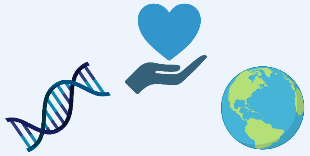
2.3 Mixtures Analyses

2.4 -Omics Analyses and Systems Biology

2.5 Toxicokinetic Modeling

2.6 Read-Across Toxicity Predictions

Environmental Health Database Mining



Training Modules

3.1 Comparative Toxicogenomics Database

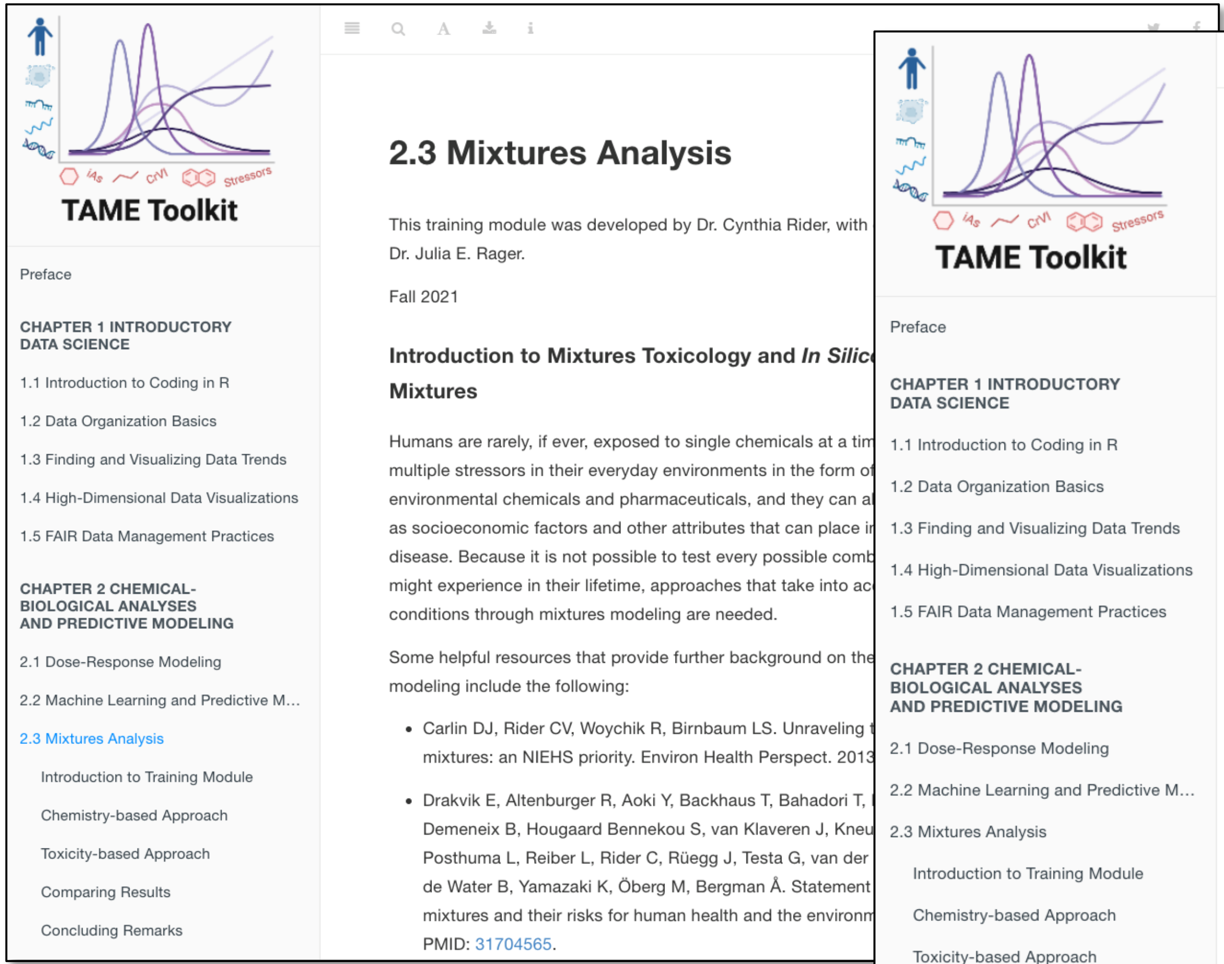
3.2 Gene Expression Omnibus

3.3 Integration across Air Quality, Mortality, and Environmental Justice Data

Content within Each Specific Module

- Introduction to the Topic
- Introduction to the Training Module
- List of Training Module's ***Environmental Health Questions***
- Script Preparations
- Scripted Analysis with Applications-based Environmental Health Questions Sprinkled throughout to Keep Participants Engaged
- Concluding Remarks and Additional Resources

Screenshots in the Mixtures Sufficient Similarity Module



TAME Toolkit

2.3 Mixtures Analysis

This training module was developed by Dr. Cynthia Rider, with ... Dr. Julia E. Rager.

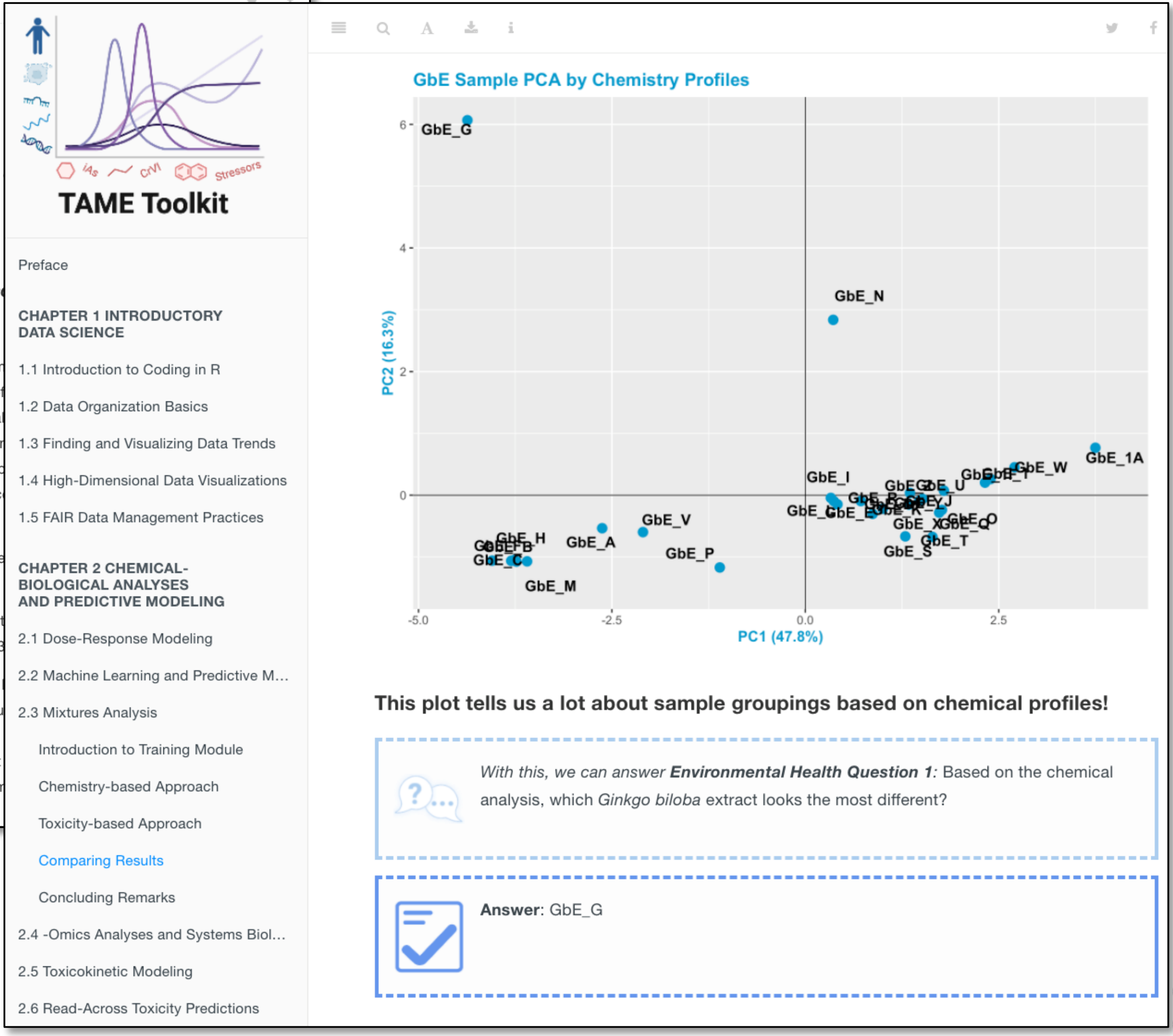
Fall 2021

Introduction to Mixtures Toxicology and *In Silic*... Mixtures

Humans are rarely, if ever, exposed to single chemicals at a tim... multiple stressors in their everyday environments in the form of... environmental chemicals and pharmaceuticals, and they can al... as socioeconomic factors and other attributes that can place i... disease. Because it is not possible to test every possible comb... might experience in their lifetime, approaches that take into ac... conditions through mixtures modeling are needed.

Some helpful resources that provide further background on the... modeling include the following:

- Carlin DJ, Rider CV, Woychik R, Birnbaum LS. Unraveling t... mixtures: an NIEHS priority. Environ Health Perspect. 2013...
- Drakvik E, Altenburger R, Aoki Y, Backhaus T, Bahadori T, ... Demeneix B, Hougaard Bennekou S, van Klaveren J, Kneu... Posthuma L, Reiber L, Rider C, Rüegg J, Testa G, van der... de Water B, Yamazaki K, Öberg M, Bergman Å. Statement... mixtures and their risks for human health and the environm... PMID: 31704565.



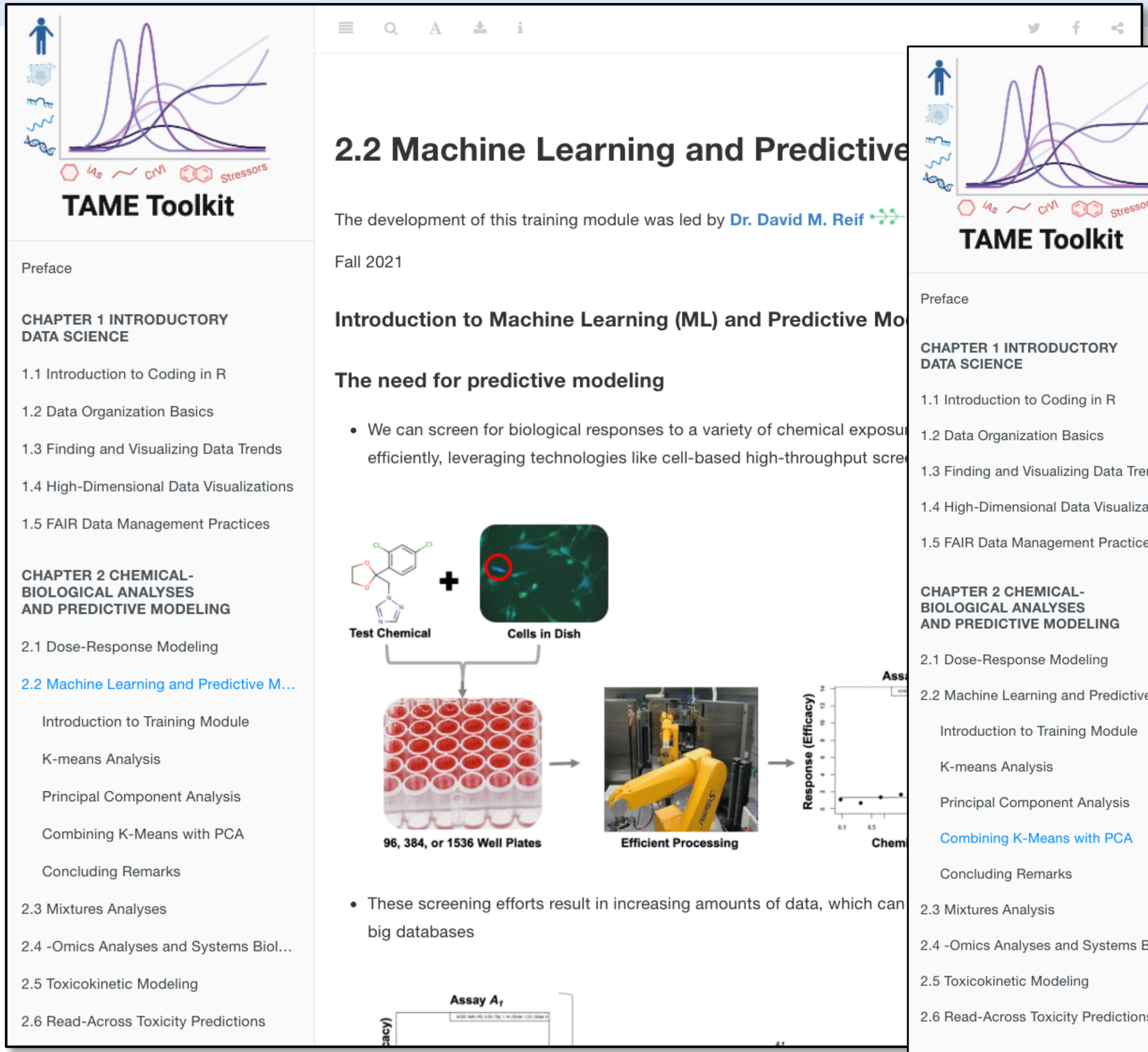
TAME Toolkit

This plot tells us a lot about sample groupings based on chemical profiles!

*With this, we can answer **Environmental Health Question 1**: Based on the chemical analysis, which Ginkgo biloba extract looks the most different?*

Answer: GbE_G

Screenshots in a Machine Learning Module


TAME Toolkit

iAs CrVI Stressors

2.2 Machine Learning and Predictive

The development of this training module was led by **Dr. David M. Reif**

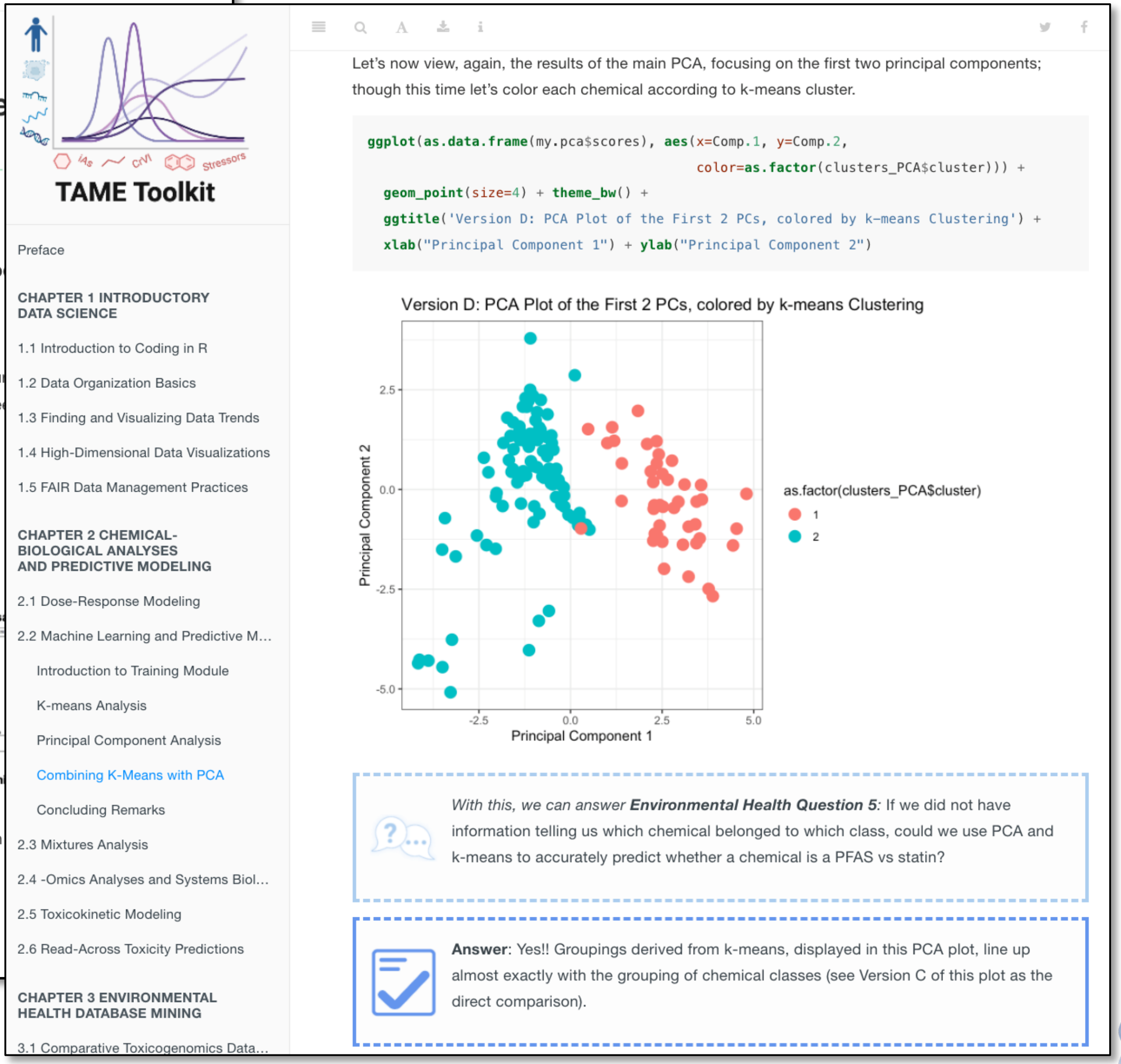
Fall 2021

Introduction to Machine Learning (ML) and Predictive Mo

The need for predictive modeling

- We can screen for biological responses to a variety of chemical exposu efficiently, leveraging technologies like cell-based high-throughput scre

- These screening efforts result in increasing amounts of data, which can big databases


TAME Toolkit

Let's now view, again, the results of the main PCA, focusing on the first two principal components; though this time let's color each chemical according to k-means cluster.

```
ggplot(as.data.frame(my.pca$scores), aes(x=Comp.1, y=Comp.2,
                                         color=as.factor(clusters_PCA$cluster))) +
  geom_point(size=4) + theme_bw() +
  ggtitle('Version D: PCA Plot of the First 2 PCs, colored by k-means Clustering') +
  xlab("Principal Component 1") + ylab("Principal Component 2")
```


With this, we can answer ***Environmental Health Question 5****:* If we did not have information telling us which chemical belonged to which class, could we use PCA and k-means to accurately predict whether a chemical is a PFAS vs statin?

Answer: Yes!! Groupings derived from k-means, displayed in this PCA plot, line up almost exactly with the grouping of chemical classes (see Version C of this plot as the direct comparison).

Screenshots in the -Omics Analyses Module

TAME Toolkit

2.4 -Omics Analyses and System

This training module was developed by Lauren Koval, Dr. Kyle Roell,

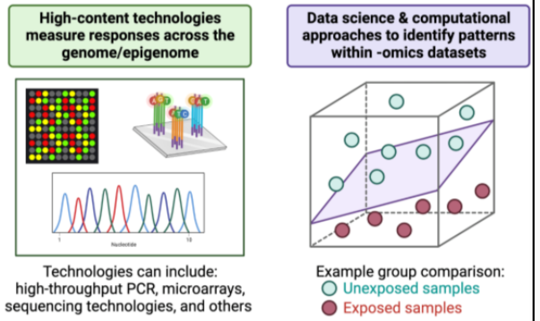
Fall 2021

The Field of "-Omics"

The field of "-omics" has rapidly evolved since its inception in the mi obtained through sequencing of the human genome (see the Human advent of high-content technologies. High-content technologies hav assessment of genome-wide, or 'omics'-based, endpoints.

Traditional molecular biology techniques typically evaluate the functi products. Omics-based methods, on the other hand, utilize non-targ genes or gene products in a given environmental/biological sample. allow for the unbiased investigation of potentially unknown or unders involved in regulating cell health and disease. These molecular profil altered in response to toxicant exposures and/or during disease initia

To further understand the molecular consequences of -omics-based overlaid onto molecular networks to uncover biological pathways an perturbed at the systems biology level. An overview of these general content technologies and ending of systems biology, is provided in t BioRender.com).



TAME Toolkit

Preface

CHAPTER 1 INTRODUCTORY DATA SCIENCE

1.1 Introduction to Coding in R

1.2 Data Organization Basics

1.3 Finding and Visualizing Data Trends

1.4 High-Dimensional Data Visualizations

1.5 FAIR Data Management Practices

CHAPTER 2 CHEMICAL-BIOLOGICAL ANALYSES AND PREDICTIVE MODELING

2.1 Dose-Response Modeling

2.2 Machine Learning and Predictive M...

2.3 Mixtures Analysis

2.4 -Omics Analyses and Systems Biol...

The Field of "-Omics"

Transcriptomics

Introduction to Training Module

Transcriptomics Data QA/QC

Statistical Analysis of Gene Expressi...

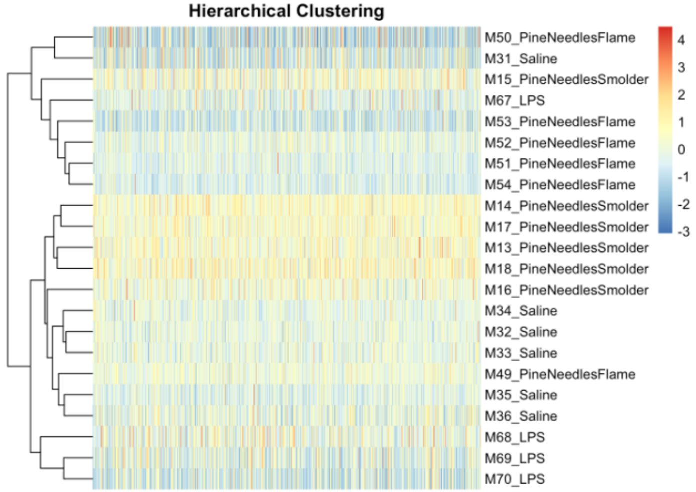
MA Plots

Volcano Plots

Pathway Enrichment Analysis

Concluding Remarks


```
pheatmap(scale(countdata_for_clustering), main="Hierarchical Clustering",
         cluster_rows=TRUE, cluster_cols = FALSE,
         fontsize_col = 7, treeheight_row = 60, show_colnames = FALSE)
```



Like the PCA findings, heirarchical clustering demonstrated an overall lack of potential sample outliers because there were no obvious sample(s) that grouped separately from the rest along the clustering dendograms.
Therefore, *neither approach points to outliers that should be removed in this analysis.*

With this, we can answer ***Environmental Health Question 2****:* When preparing transcriptomics data for statistical analyses, what are three common data filtering steps that are completed during the data QA/QC process?

Answer: (1) Background filter to remove genes that are universally lowly expressed; (2) Sample filter to remove samples that may be not have any detectable mRNA; (3) Sample outlier filter to remove samples with underlying data distributions outside of the overall, collective dataset.

TAME Toolkit Dissemination

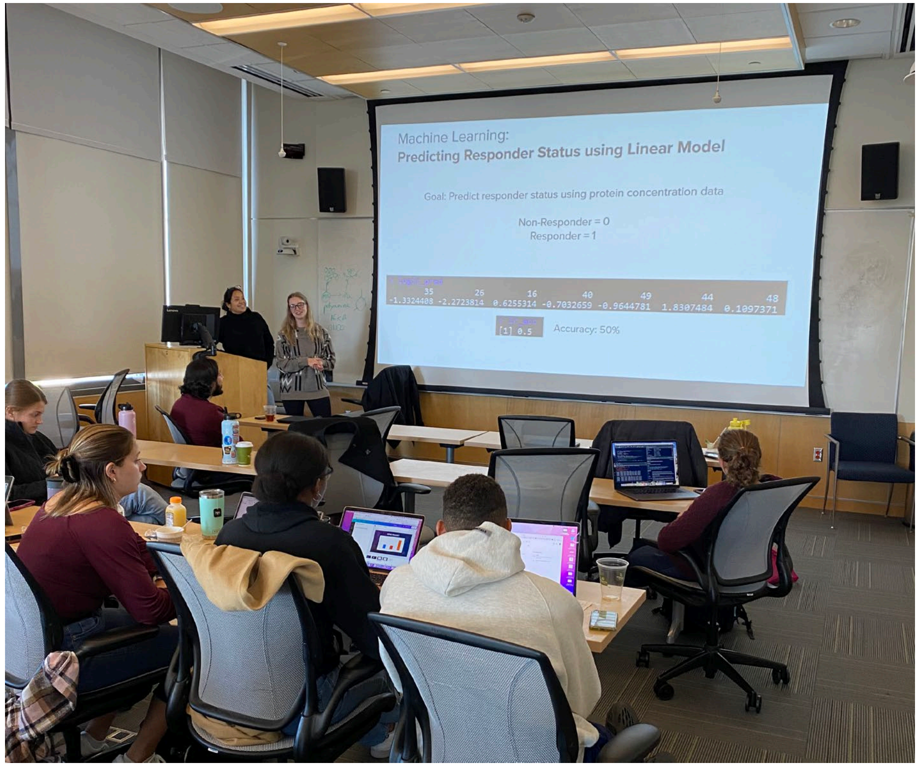
Course Materials

- Much of this training content was tested through a new course at UNC, ***ENVR 730: Computational Toxicology and Exposure Science***
- Included 17 graduate students from:
  - Department of Environmental Sciences and Engineering (ENVR)
  - Curriculum of Toxicology and Environmental Medicine (CiTEM)
- Received the **Teaching Innovation Award** from the UNC Gillings School of Global Public Health, which is student-nominated!
- We are teaching this course again this Fall

Other Dissemination Efforts

High dimensional data visualization training workshop (virtual, Spring 2022)

PRogramming for Environmental HEalth And Toxicology (PREHEAT) Retreat



International Engagement

Viewers have accessed the TAME toolkit world-wide

Google analytics (Sept 2022-current):

Global views across 91 countries:

5,300 total site views

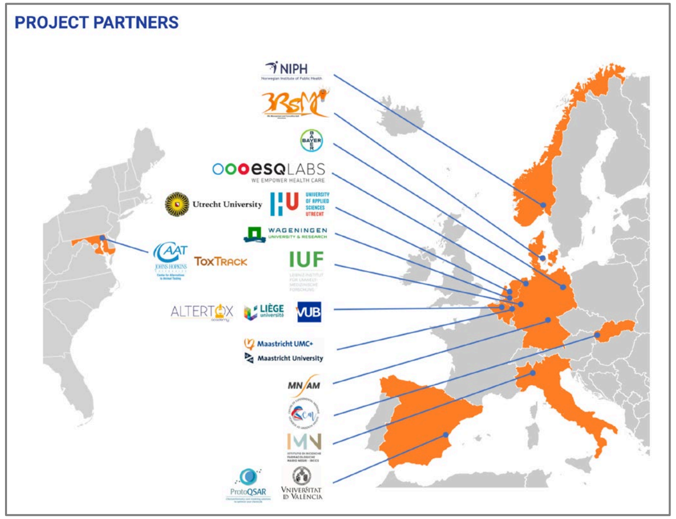
1,441 full site users

Number of users per country with highest usage:

COUNTRY	USERS
United States	600
Germany	62
Canada	60
India	59
Japan	44
China	39
France	38

International collaborators

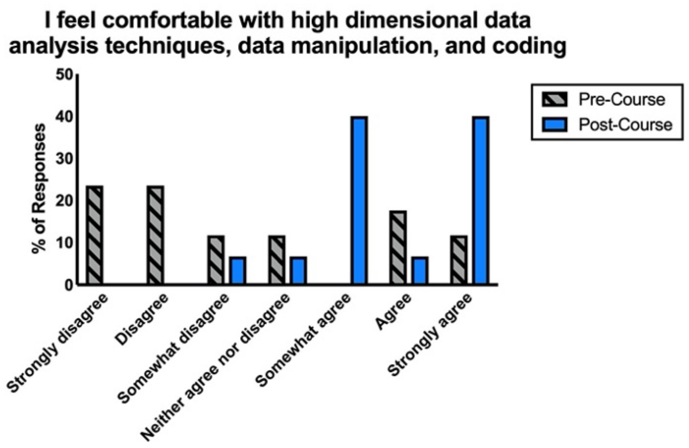
For example, the **European Union's ONTOX** project leaders are incorporating TAME materials into their training courses on data analysis and machine learning in toxicology

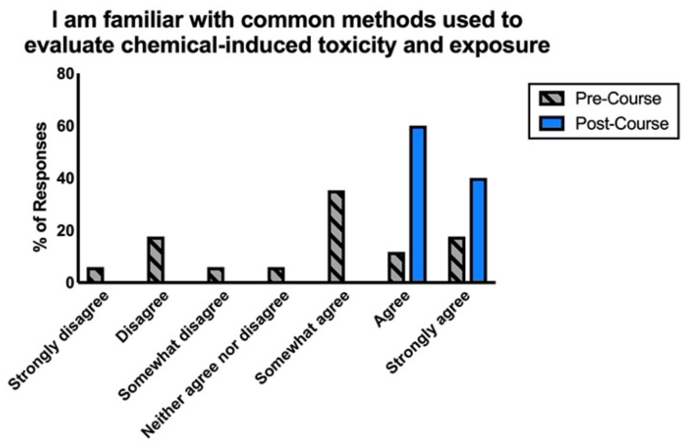


https://ontox-project.eu/consortium/

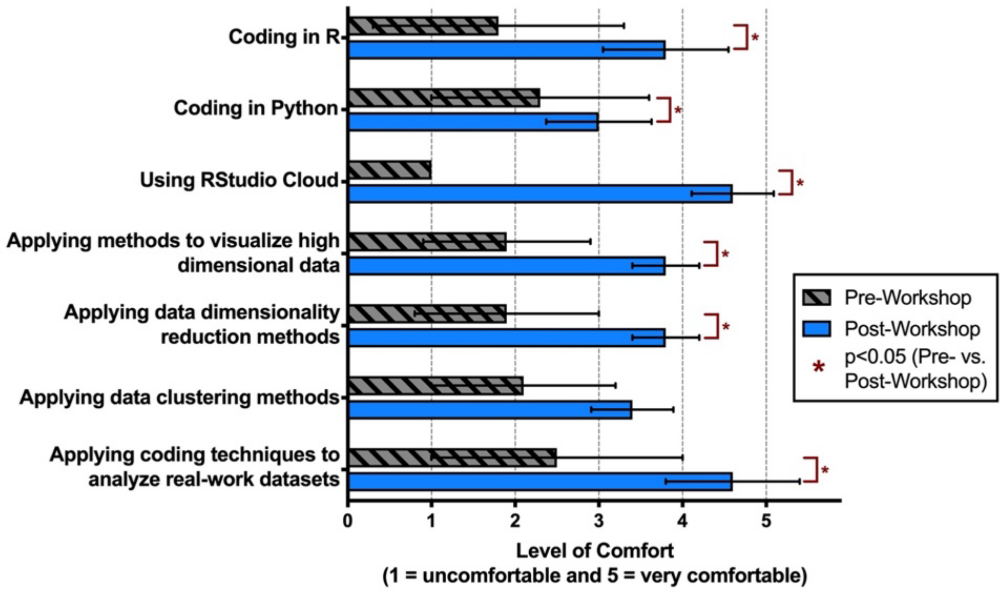
TAME Effectiveness and Feedback

Course Evaluation:





Virtual Workshop Evaluation:



Consistent Feedback across All Dissemination Efforts

- **Participants want more!** More training time, more modules
- Training at the intro-level, mid-level, and advanced-level
- More basics, including best practices for wet lab data organization and sharing
- We are currently in the process of expanding these training materials…

TAME 2.0 Is Underway!

Expanding and reorganizing content into 6 chapters:

1. Introduction to Data Science
2. Introduction to Coding in R
3. Basics of Data Analysis and Visualizations
4. Machine Learning and Artificial Intelligence
5. Applications in Toxicology and Exposure Science
6. Environmental Health Database Mining

Hoping to launch ~December 2023

Collaborations and Funding

Rager lab
Celeste Carberry (PhD cand)
Elise Hickman (post-doc)
Lauren Koval (PhD cand)
Elena McDermott (MSPH)
Alexis Payton (data analyst)
Sarah Miller (PhD cand)
Kai Malone (undergrad)
Raquel Winker (undergrad)

UNC IEHS / SRP
Rebecca Fry
Audrey Bousquet (res scientist)
Lauren Eaves (res associate)
Hadley Hartwell (lab manager)
Kyle Roell (data analyst)

NIH / NIEHS
Nicole Niehoff
Matthew Wheeler
Alexis White

UNC
Stephanie Engel
Ilona Jaspers
Alex Keil
Kun Lu
Tracy Manuck
Meghan Rebuli
Gregory Smith

RTI
Rebecca Boyles
Shaun McCullough

NIH / DTT
Scott Auerbach
Stephen Ferguson
Kyle Messier
David Reif
Cynthia Rider

US EPA
Kathie Dionisio
M Ian Gilmour
Kristin Isaacs
Yong Ho Kim
Grace Patlewicz
Katie Paul-Friedman
Caroline Ring
John Wambaugh
Cavin Ward-Caviness

Trainees in italics

This work was supported by:
NIH
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
IEHS Institute for Environmental Health Solutions
CiTEM
UNC Chapel Hill SUPERFUND Research Program
CEMALB

Kyle Roell’s Walk Through TAME

- Dr. Roell, UNC-SRP lead data analyst