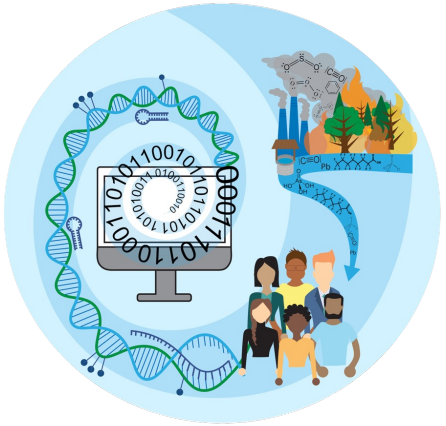
Interindividual Variability Assessment Through Application of Machine Learning with In Vitro Molecular Profiles to Understand Key Mechanisms of Emerging Inhaled Toxicants



Elise Hickman

Postdoctoral Fellow, Rager Lab

The University of North Carolina at Chapel Hill

Background:
Interindividual Variability in Risk Assessment

- Human health risks are known to vary across and within populations.

- Current questions/challenges in risk assessment include:

    1. How can we improve assessment of human interindividual variability?

    2. How can we improving linkages between exposures that include multiple stressors and disease outcomes across the full range of human responses?

    3. How can we determine uncertainty factors that are applicable to specific endpoints and exposures and that capture interindividual variability?

How can machine learning help us understand interindividual variability?

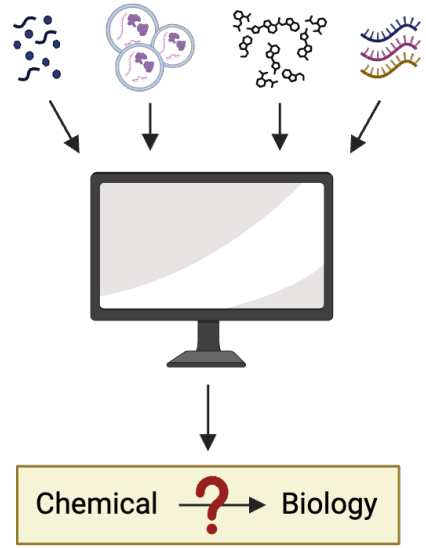
Rusyn, Chiu, Wright 2022; Regulatory Toxicology and Pharmacology; DOI: 10.1016/j.yrtph.2022.105197

Big (and Smaller!) Data

Technological advances have made measuring molecular signatures in experimental samples more feasible and affordable.

Pros:

- Increased accessibility of measuring a wide range of molecular signatures
- Opportunity for broader investigation of the effects of toxicants
- Higher sensitivity in capturing molecular signatures
- Ability to obtain more data from a single sample



Challenges:

- Sufficiently powering studies
- Distilling meaningful biological conclusions AND communicating them clearly
- Data science training

Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.

Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.

Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?

What are e-cigarettes?

E-cigarettes heat and aerosolize an e-liquid, allowing users to inhale nicotine and other chemicals.

E-cigarettes were originally touted as a “safer” alternative to cigarettes but are used by both former cigarette smokers and nonsmokers.

E-liquids typically contain:

- Nicotine or Nicotine Salts, 0-7% (0-70 mg/mL)
- Flavoring Chemicals
- Propylene Glycol (throat hit)
- Vegetable Glycerin (sweetness, cloud)

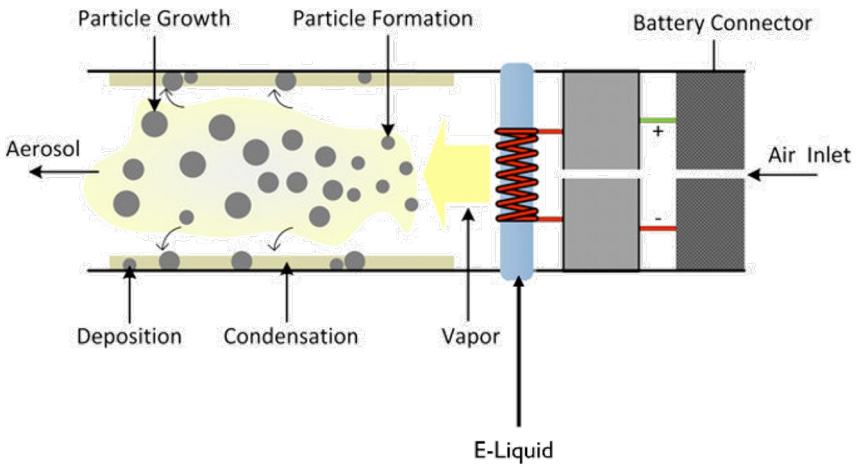
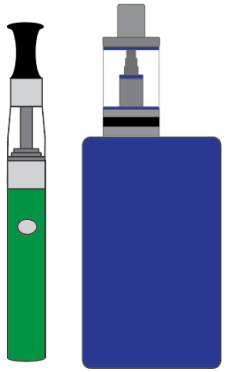


Image: Talih 2016 Aerosol Science and Technology

E-Cigarette Device Evolution

Constant evolution of e-cigarette devices is a major challenge in the field of e-cigarette toxicology, particularly with popular devices such as JUUL and disposables.

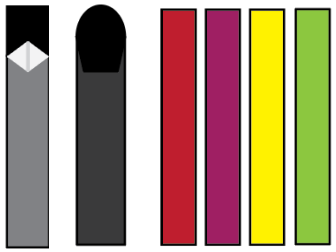
Previous (3rd) Generation E-Cigs



Freebase Nicotine
(pH ~8)

User control over settings, can be high power

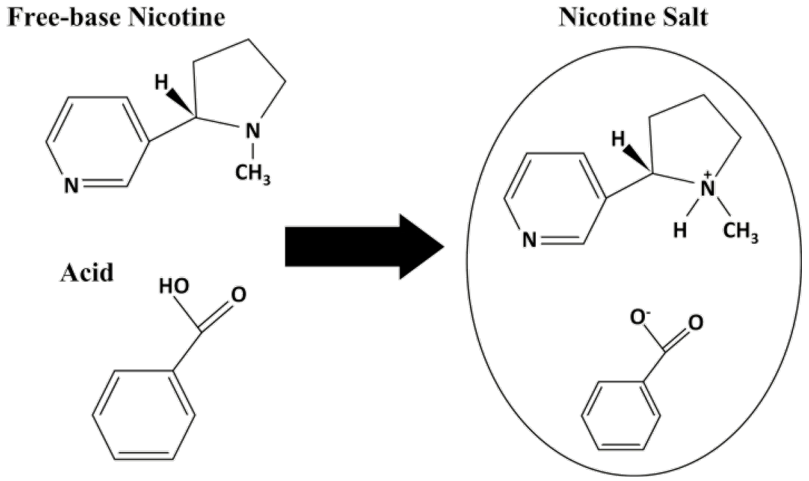
Pod and Disposable E-Cigs (4th Gen)



Nicotine Salts
(pH ~4-6)

Fewer settings, lower power

Puffing patterns? User population?



e.g. lactic, benzoic, and levulinic acids

What biomarkers are altered in 4th generation e-cigarette users?

Image: Harvanko et al 2020 Nicotine Tob Res

Study Design

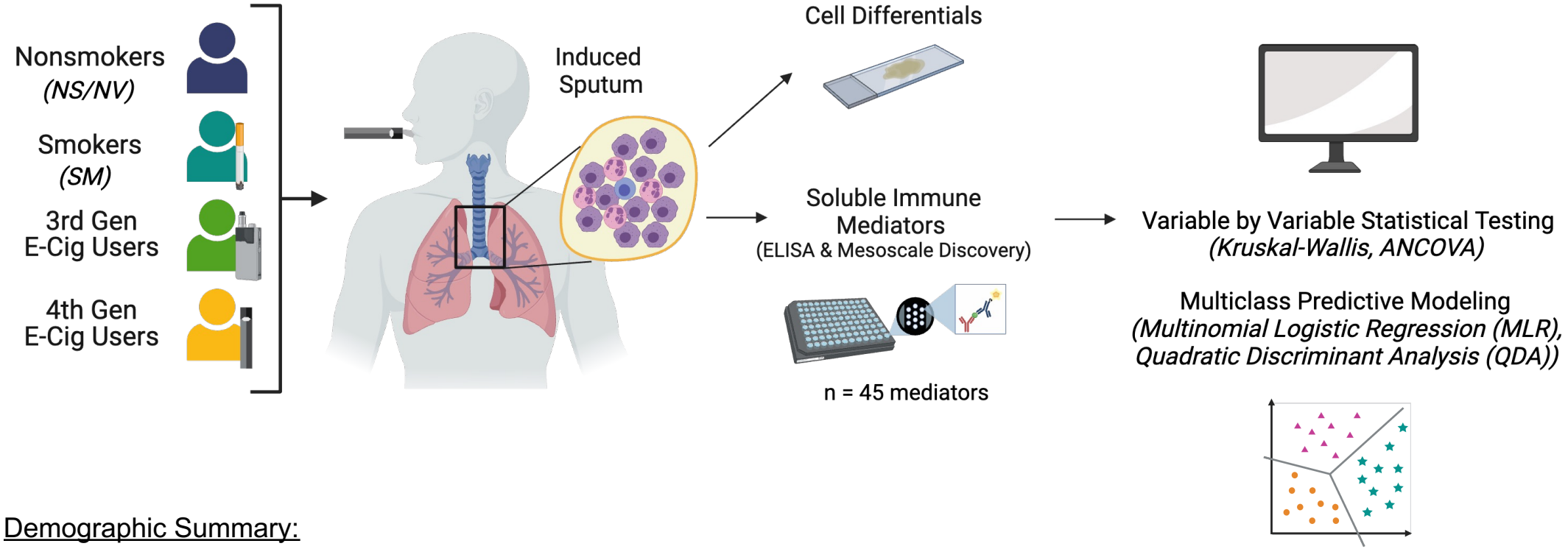

Dr. Ilona Jaspers


Dr. Neil Alexis


Heather Wells


Dr. Julia Rager


Alexis Payton



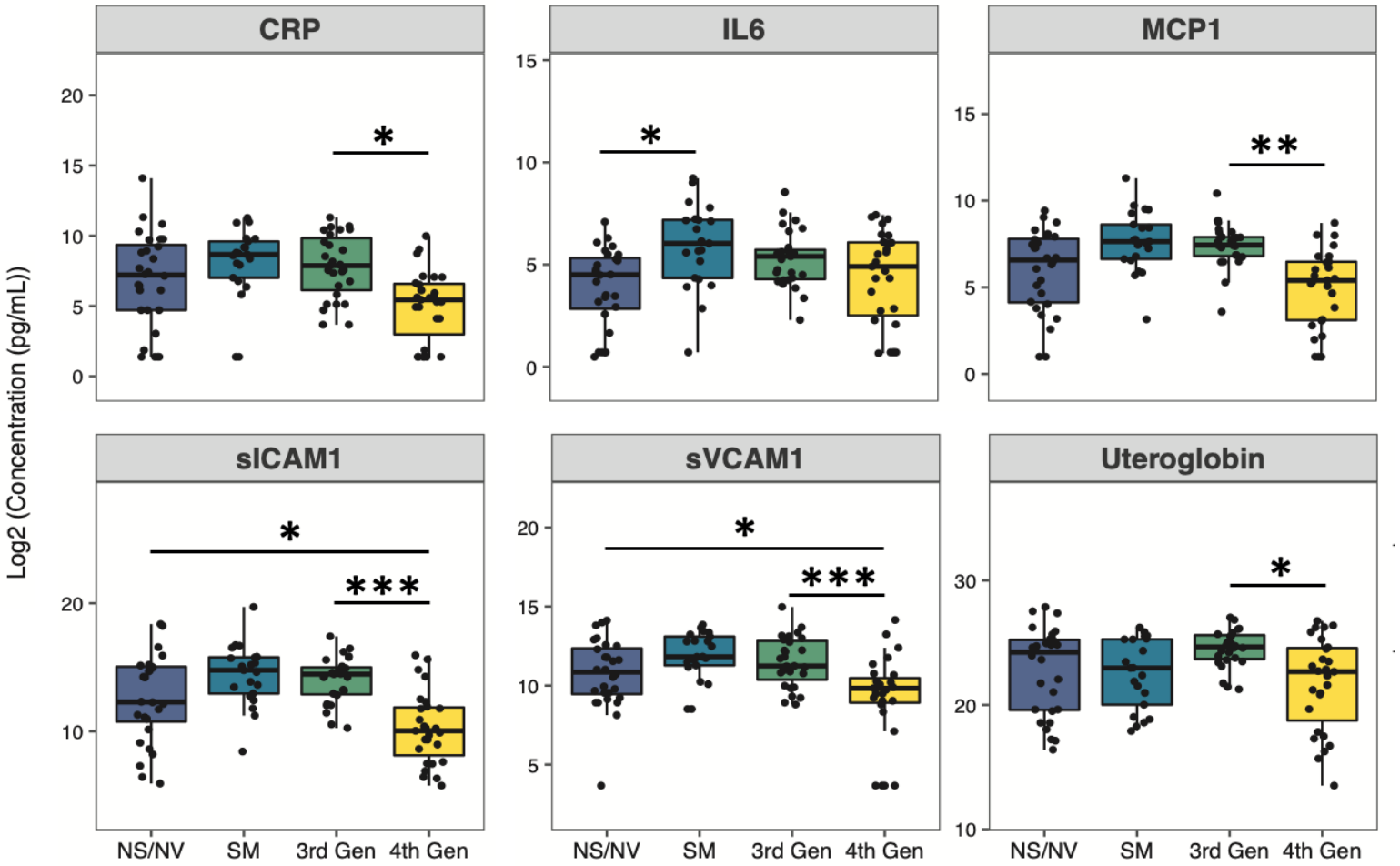
Demographic Summary:

- n = 21-28 participants per group
- 4th generation e-cigarette users were significantly younger
- Each group had a mixture of male and female participants, but ratio was not always even

Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine

Soluble Mediator Expression is Significantly Decreased in 4th Generation E-Cig Users

n = 12 mediators significant for exposure group variable



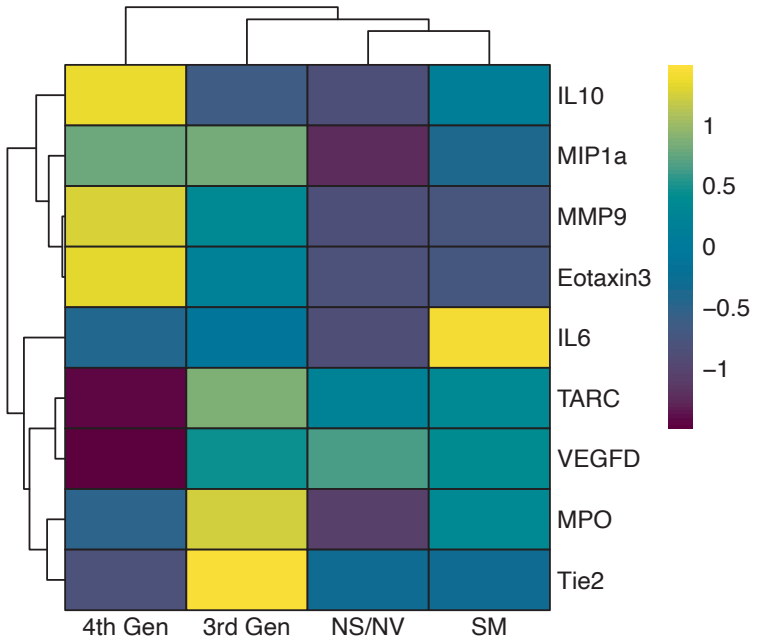
Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine

*Soluble mediators that were significantly different between exposure groups after adjusting for age, sex, and race differences between exposure groups. Results are presented as mean ± standard error of log2 transformed mediator concentrations. * p < 0.05, ** p < 0.01, *** p < 0.001 using ANCOVA followed by Dunnett's (comparisons with NS/NV) and Tukey's (3rd v. 4th Gen) post-hoc tests. NS/NV = non-smoker/non-vaper, SM = smoker.*

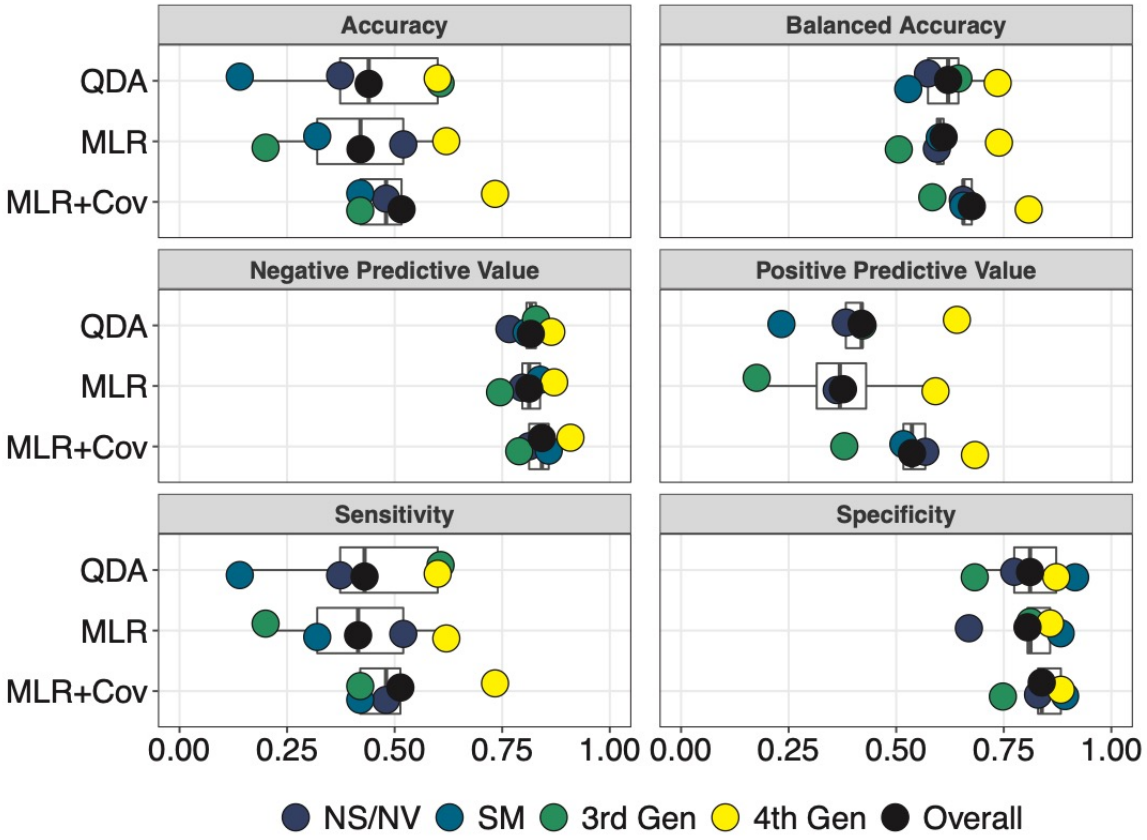
Machine learning demonstrates best separation for 4th generation e-cigarette users

Variable Selection

(Best Subsets Regression)

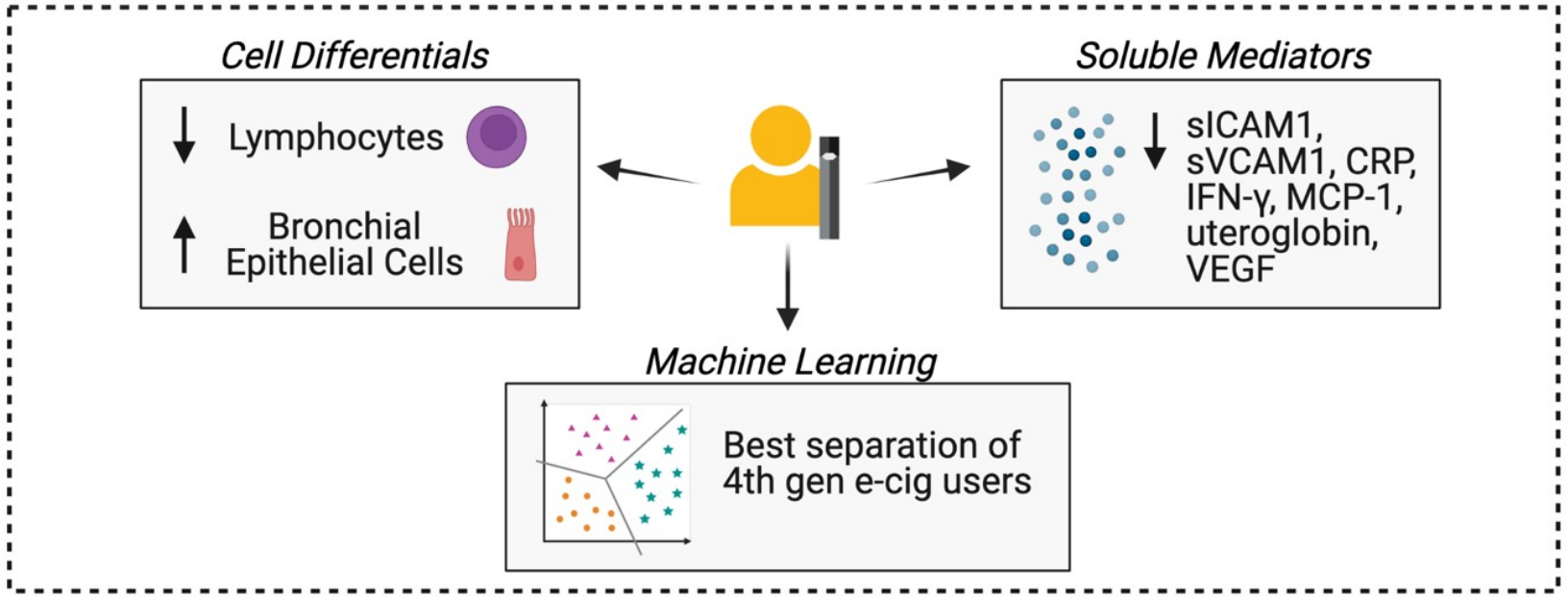


Machine Learning Model Performance



Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine

Conclusions



Suggestive of dysregulated immune homeostasis in the form of overall immune suppression in 4th generation e-cigarette users, which could result in impaired response to infection or vaccination

Observed notable interindividual variability between participants.

Hickman et al. 2022 American Journal of Respiratory and Critical Care Medicine

Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?

Example Studies

1. Are there overall differences in human respiratory protein profiles in users of different types of e-cigarette devices?

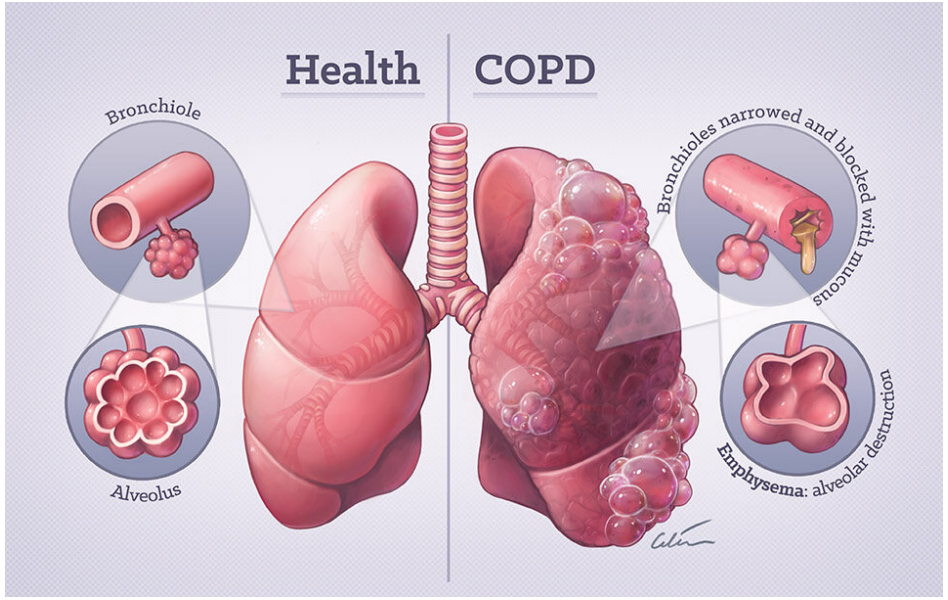
2. Are human respiratory protein profiles in e-cigarette users similar to those found in people with chronic obstructive pulmonary disease (COPD)?

Background on COPD

- Chronic obstructive pulmonary disease (COPD) is a highly prevalent, progressive condition marked by an altered airway inflammatory and immune milieu that encompasses emphysema and chronic bronchitis.

- In industrialized nations, cigarette smoking is the primary risk factor for COPD, and smoking is estimated to account for 8 in 10 COPD deaths.

- E-cig use has been associated with chronic bronchitis, increased airway proteases, inflammation, and altered immune markers in sputum, which are also found in COPD.



Do e-cig users have sputum soluble mediator profiles that resemble specific stages of COPD?

Image: https://www.vywwc.com/blog/when-to-visit-urgent-care-for-copd-recognizing-warning-signs/

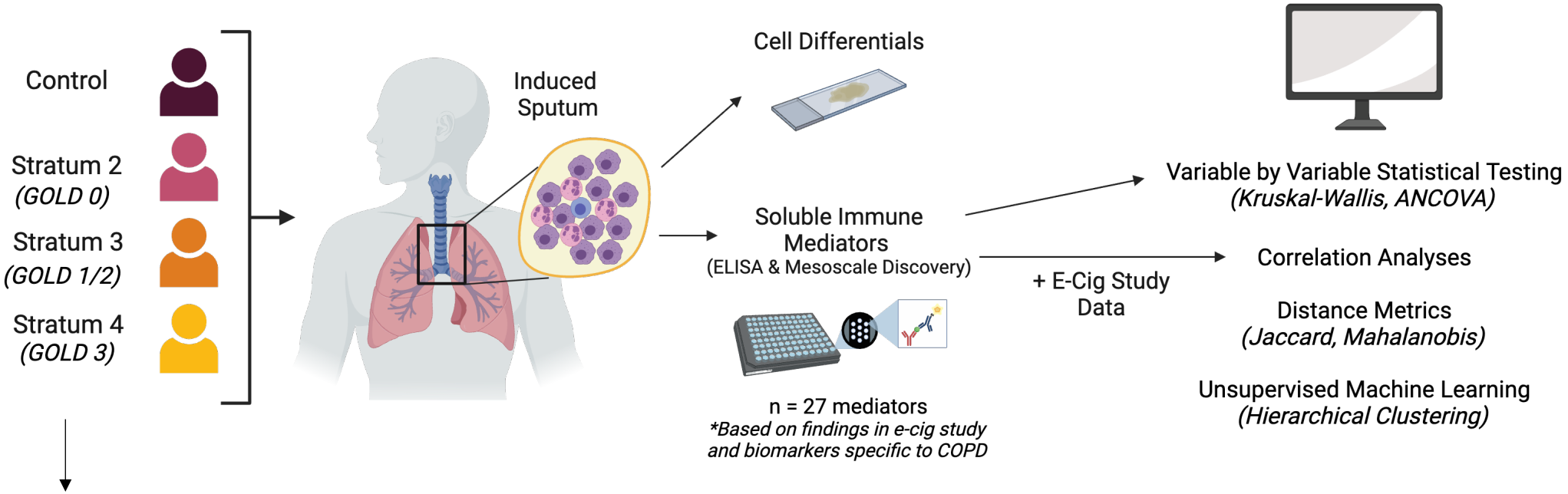
Study Design


Dr. Neil Alexis


Dr. Julia Rager

SPIROMICS
SubPopulations and InteRmediate Outcome Measures In COPD Study

SPIROMICS: **S**ub**P**opulations and **I**nte**R**mediate **O**utcome **M**easures **I**n **C**OPD **S**tudy



GOLD Stage	Description	Lung Function
0	Pre-COPD; individuals with respiratory and/or structural or physiological abnormalities without airflow obstruction	
1/2	Mild COPD (FEV1 ≥ 80% predicted), Moderate COPD (50% ≤ FEV1 < 80% predicted)	
3	Severe COPD (30% ≤ FEV1 < 50%)	

GOLD: Global Initiative for Chronic Obstructive Lung Disease

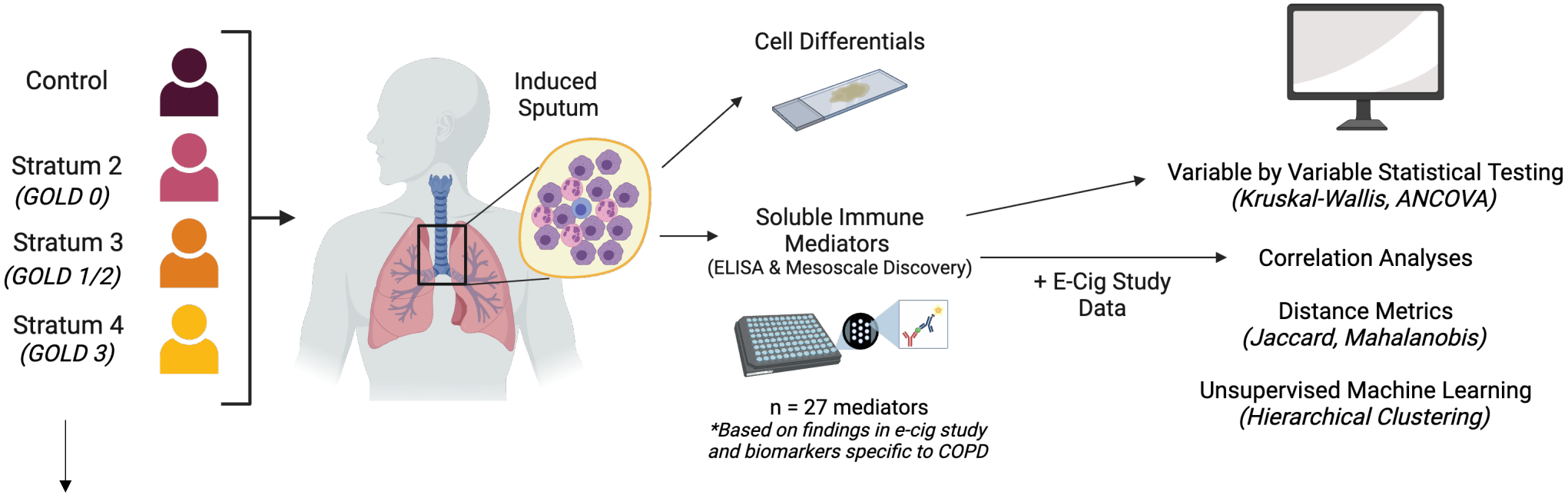
Study Design


Dr. Neil Alexis


Dr. Julia Rager

SPIROMICS
SubPopulations and InteRmediate Outcome Measures In COPD Study

SPIROMICS: **S**ub**P**opulations and **I**nte**R**mediate **O**utcome **M**easures **I**n **C**OPD **S**tudy



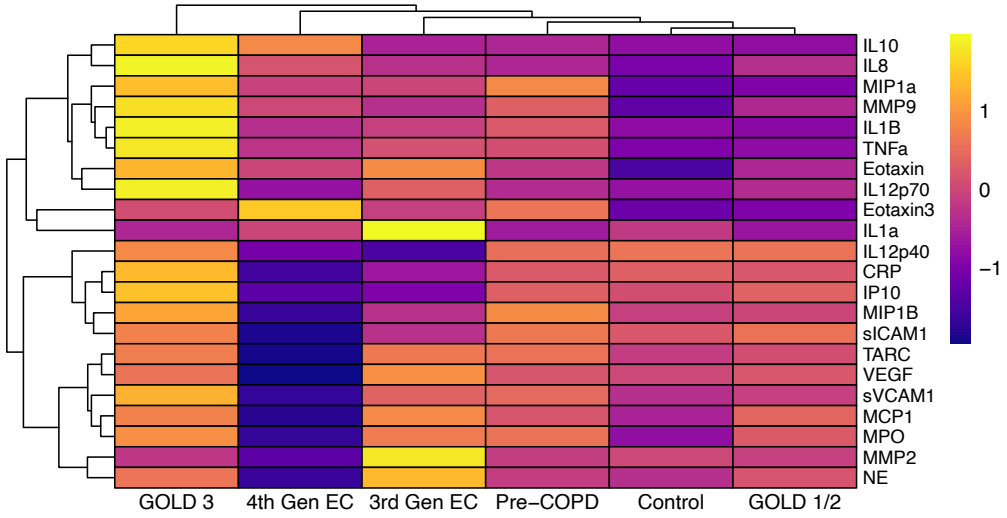
Demographic Summary:

- n = 25-29 participants per group
- Balanced male/female in each group
- Balanced current smokers vs. non-smokers in each group
- Older on average than e-cig study cohort

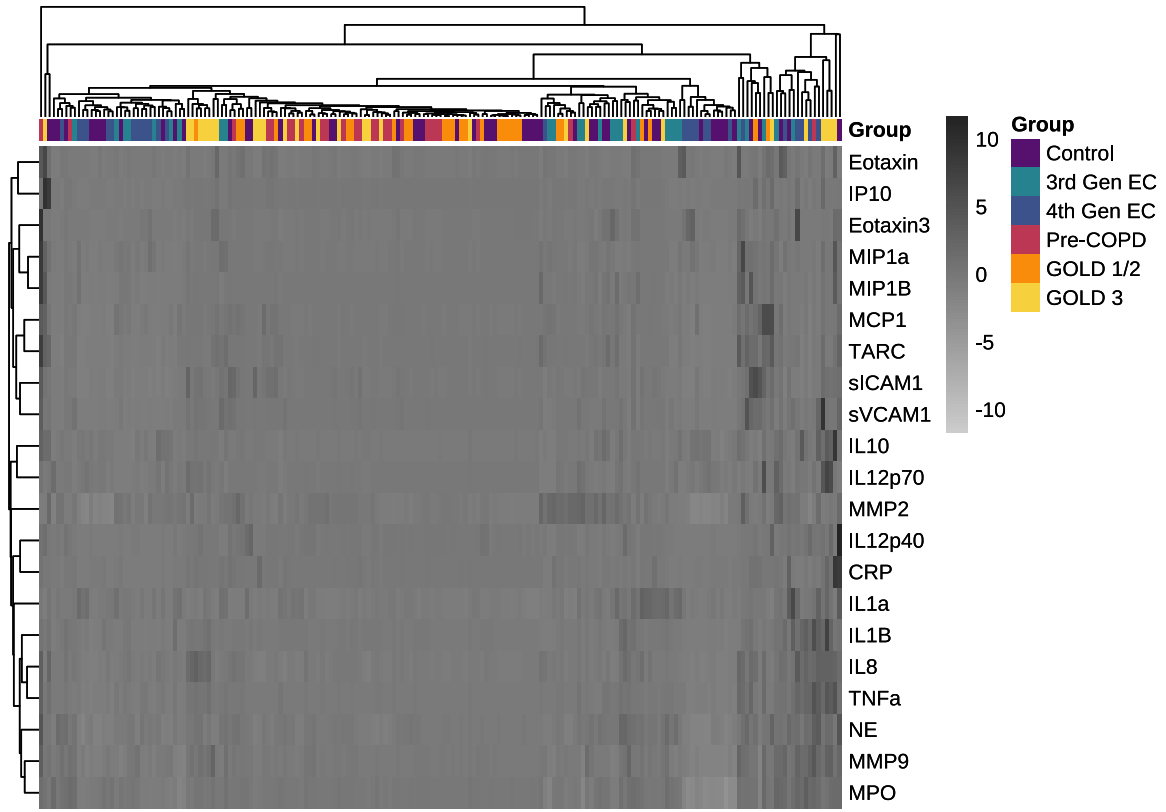
GOLD: Global Initiative for Chronic Obstructive Lung Disease

Similarity in Soluble Mediator Profiles Between Groups:
Hierarchical Clustering

All Mediators Together (Group Averages)



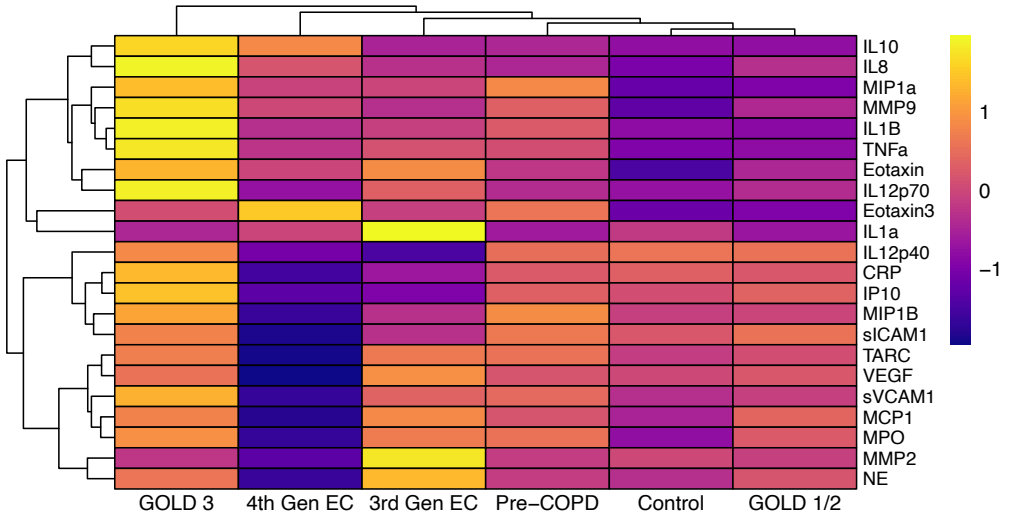
All Mediators Together (Individual Participants)



Hickman et al. manuscript submitted to Am J Resp Crit Care Med

Similarity in Soluble Mediator Profiles Between Groups:
Hierarchical Clustering

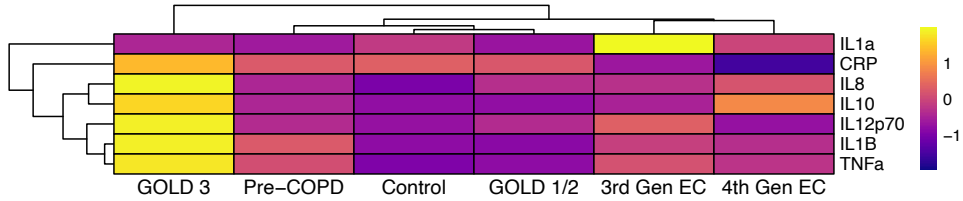
All Mediators Together (Group Averages)



Separated by Biological Function

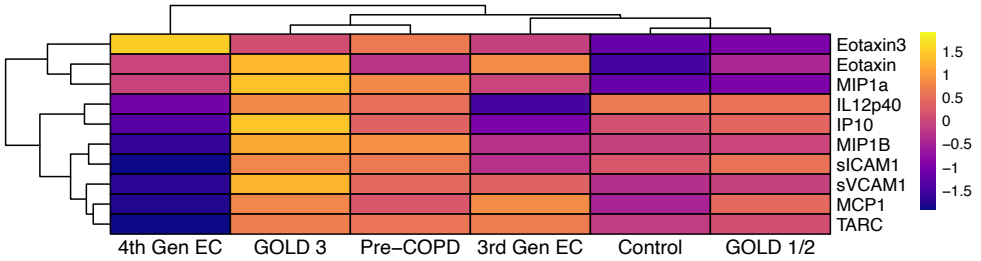
A

Inflammatory Markers



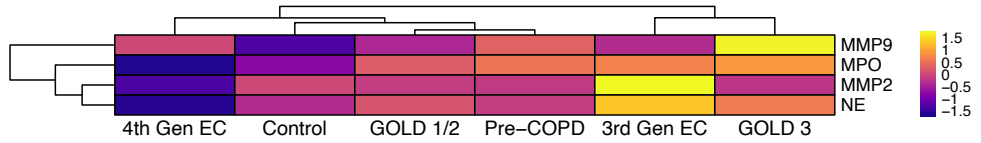
B

Chemotactic Mediators



C

Proteases & Enzymes



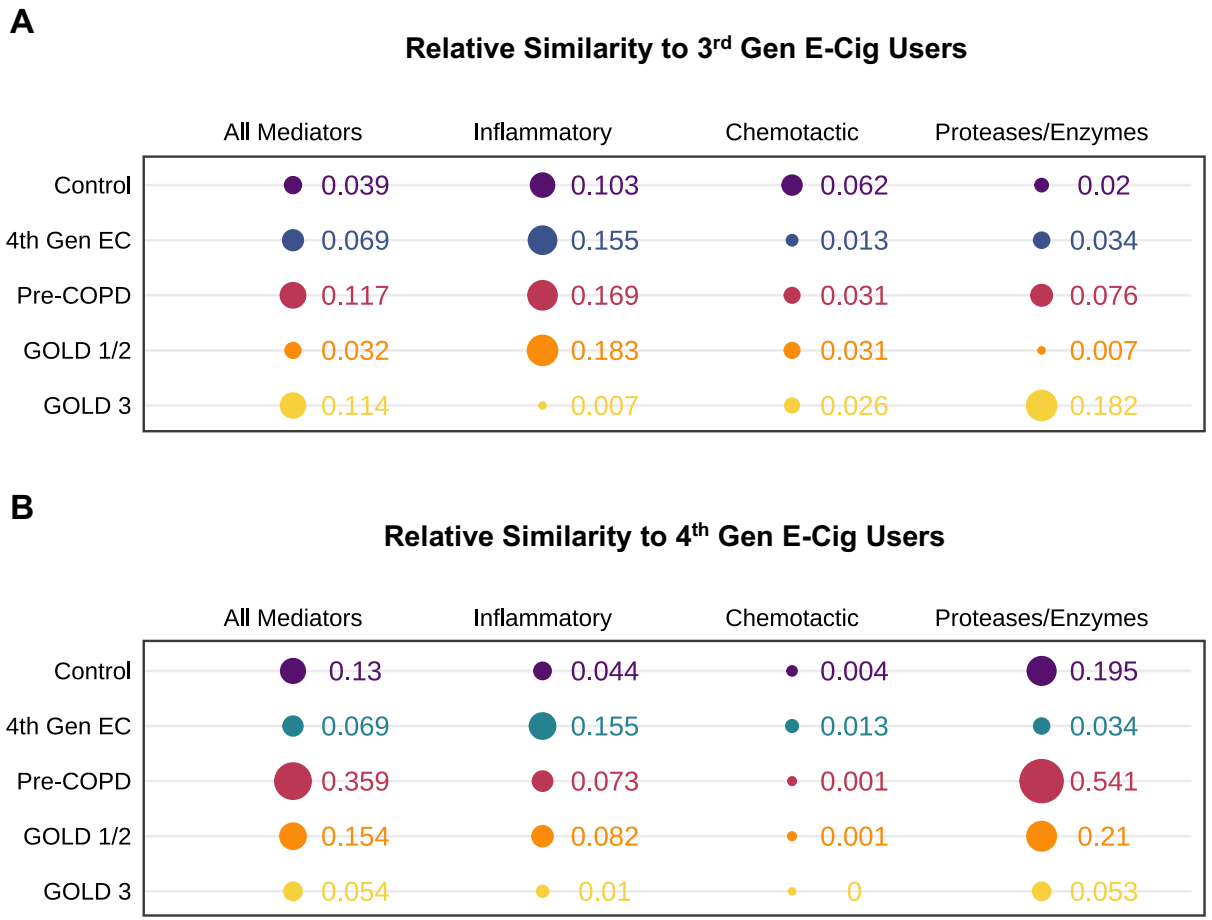
"Semi-supervised machine learning"

Hickman et al. manuscript submitted to Am J Resp Crit Care Med

Similarity in Soluble Mediator Profiles Between Groups: Mahalanobis Distance

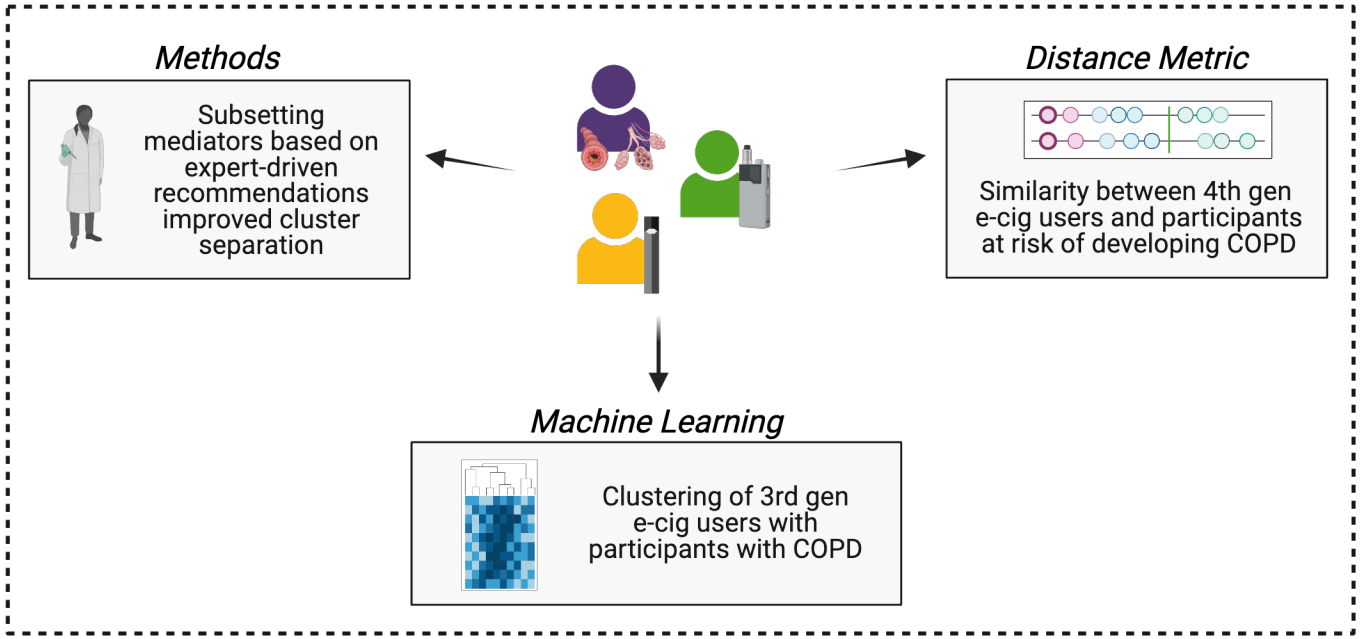
Mahalanobis distance is calculated between the multivariate mean and the datapoints after rescaling (using eigenvectors and eigenvalues) to remove covariance

Distance metrics such as Mahalanobis and Jaccard can serve as complementary approaches to machine learning.



Hickman et al. manuscript submitted to Am J Resp Crit Care Med

Conclusions



Taken together, our results demonstrate partial overlap between e-cig user and COPD soluble mediator profiles, warranting further investigation into the relationship between e-cigarette use and airway disease.

Continued to observe notable interindividual variability between participants.

Outline of Presentation

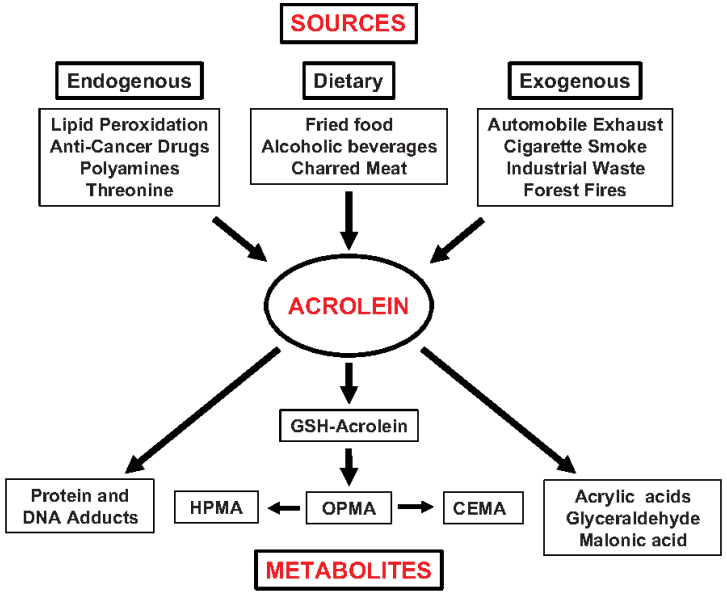
1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

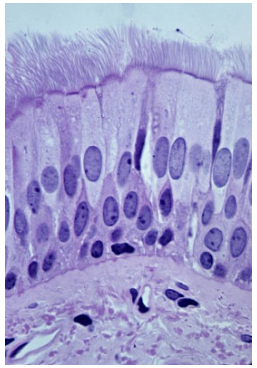
2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts. and future studies.

Background:
Acrolein & Respiratory NAMs

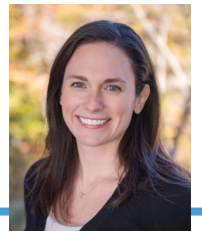
- Acrolein is a ubiquitous volatile aldehyde that is emitted from the combustion of fossil fuels, tobacco, wood, and plastic.
- Exposure to acrolein is associated with irritation throughout the respiratory tract, pulmonary edema, and dysregulation of immune responses.
- Primary human bronchial epithelial cell + fibroblast co-cultures represent sophisticated organotypic *in vitro* models that can inform interindividual variability.

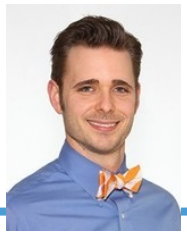


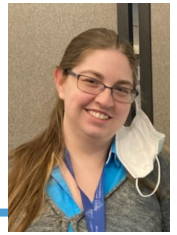


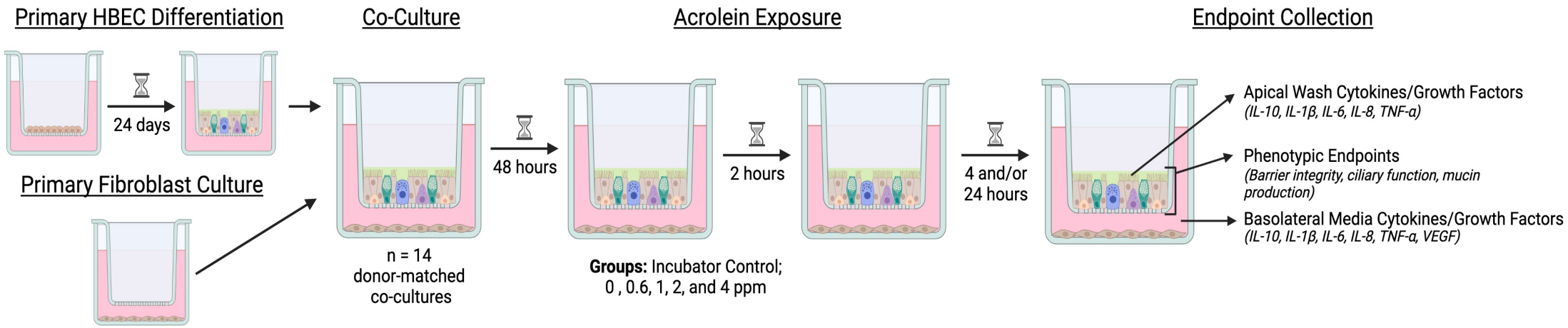
Moghe et al 2015 DOI: 10.1093/toxsci/kfu233

Cell Culture Model & Exposure


Dr. Julia Rager


Dr. Shaun McCullough
(RTI International)

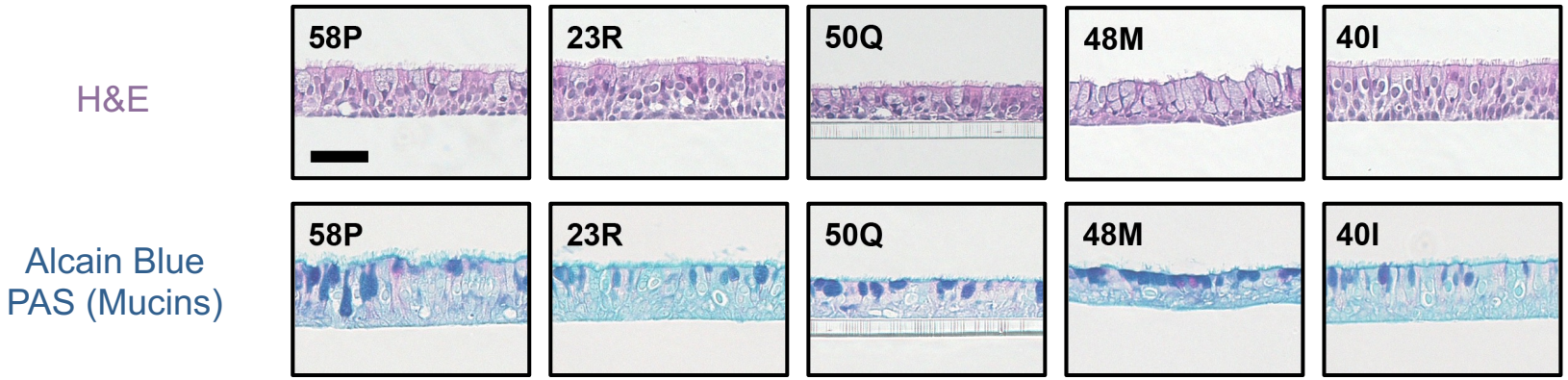

Dr. Alysha Simmons
(UNC)



Manuscript in preparation

Initial Observations

1. Significant interindividual variability between physical characteristics of pHBEC cultures.



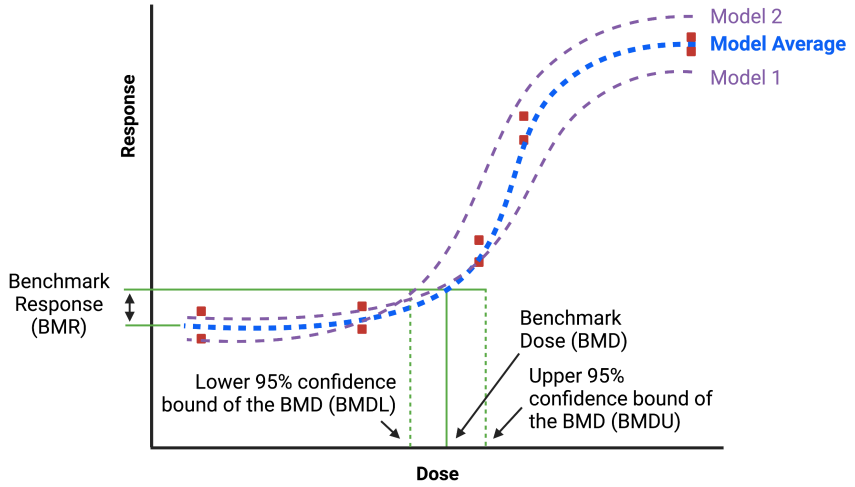
2. Significant interindividual variability in responsivity of co-culture system to acrolein exposures.
3. Significant increase in cytokine/growth factor production alongside decreased barrier integrity with higher doses of acrolein.

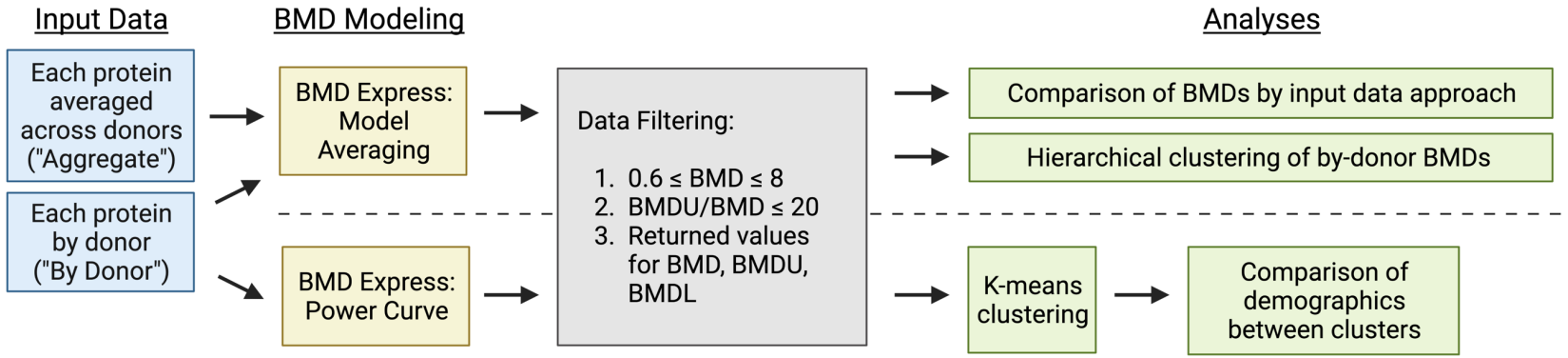
Can we leverage benchmark dose-response modeling and machine learning to assess interindividual variability in response to acrolein?

Manuscript in preparation

Computational Modeling

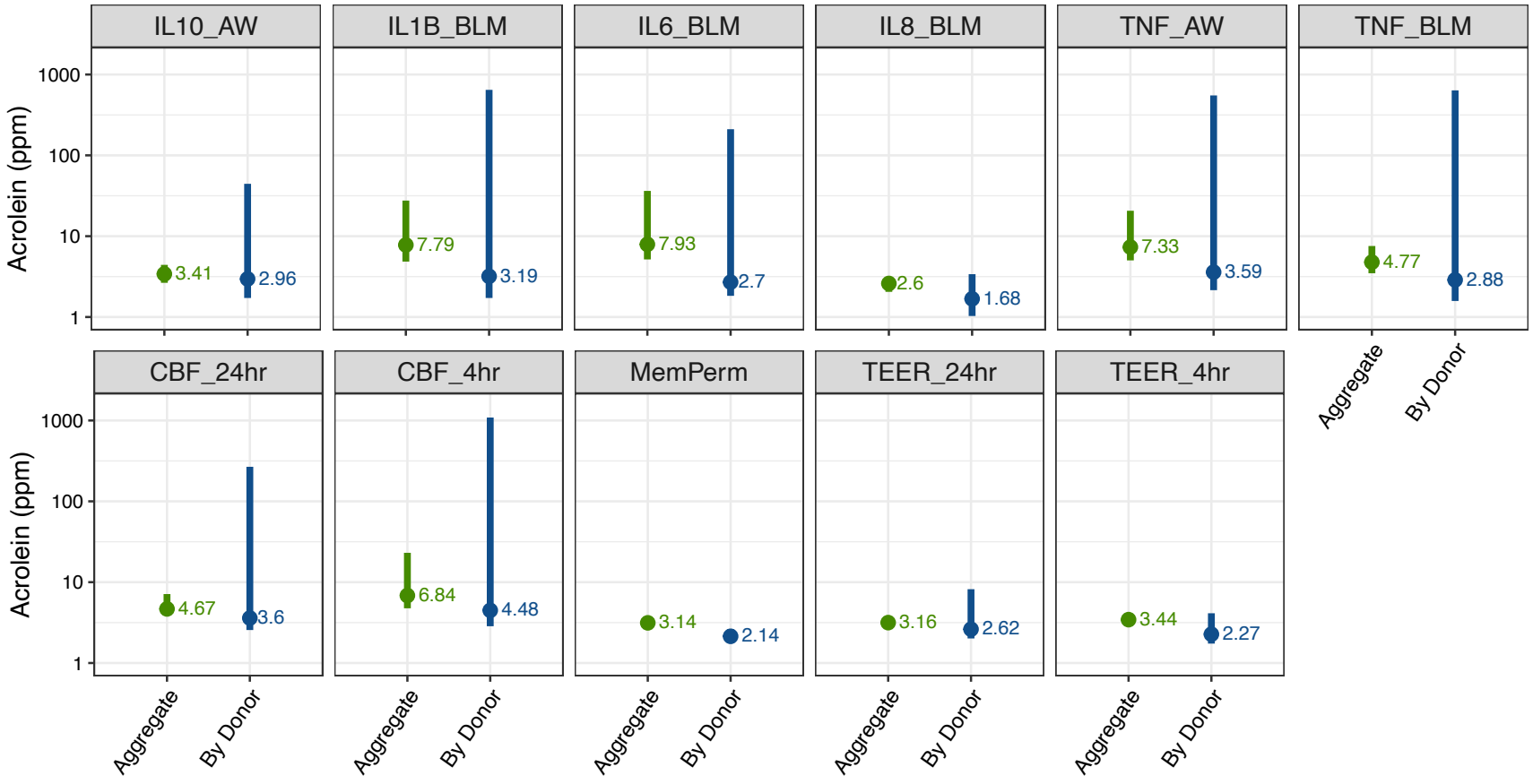
Benchmark dose-response modeling is an established tool to inform human health risk calculations that can leverage both phenotypic and molecular-level response signatures.



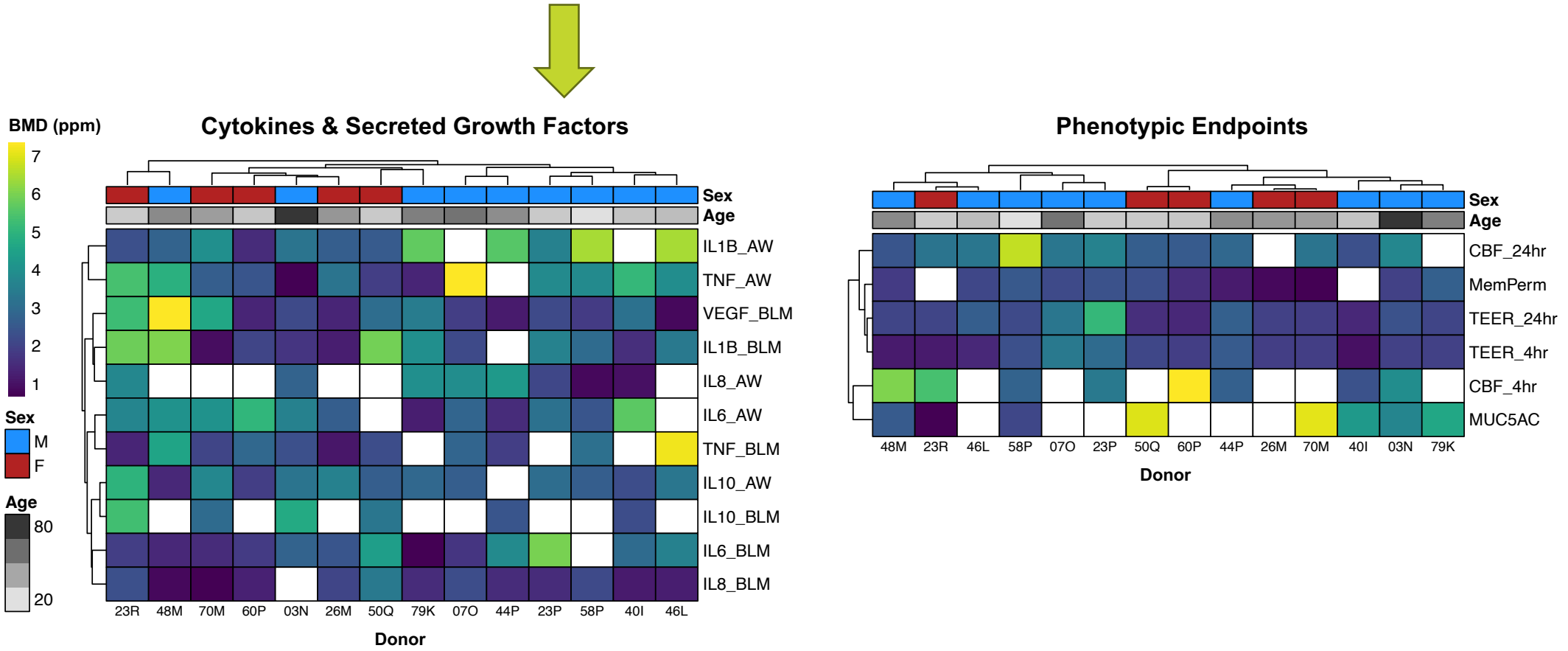


Manuscript in preparation

BMDs Were Lower and More Variable When Analyzing Trends on a Per-Donor Basis

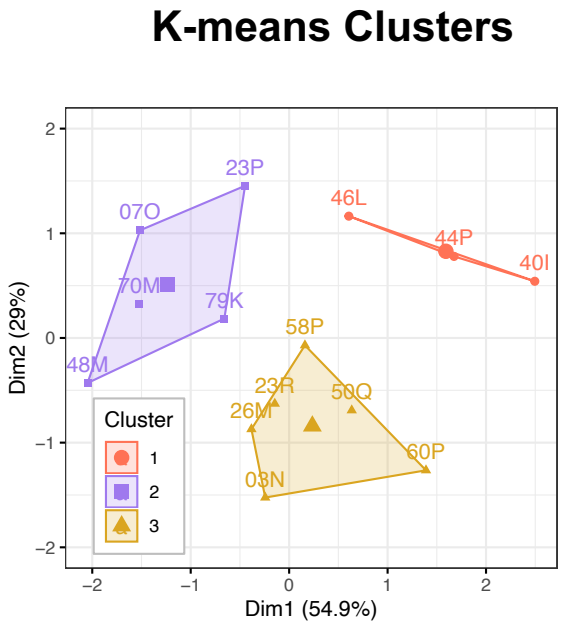


Benchmark Doses Vary by Donor and Cluster by Sex for Cytokines and Secreted Growth Factors



Manuscript in preparation

Potential Sex-Based Differences in BMD Model Parameters Were Identified Using K-Means Clustering

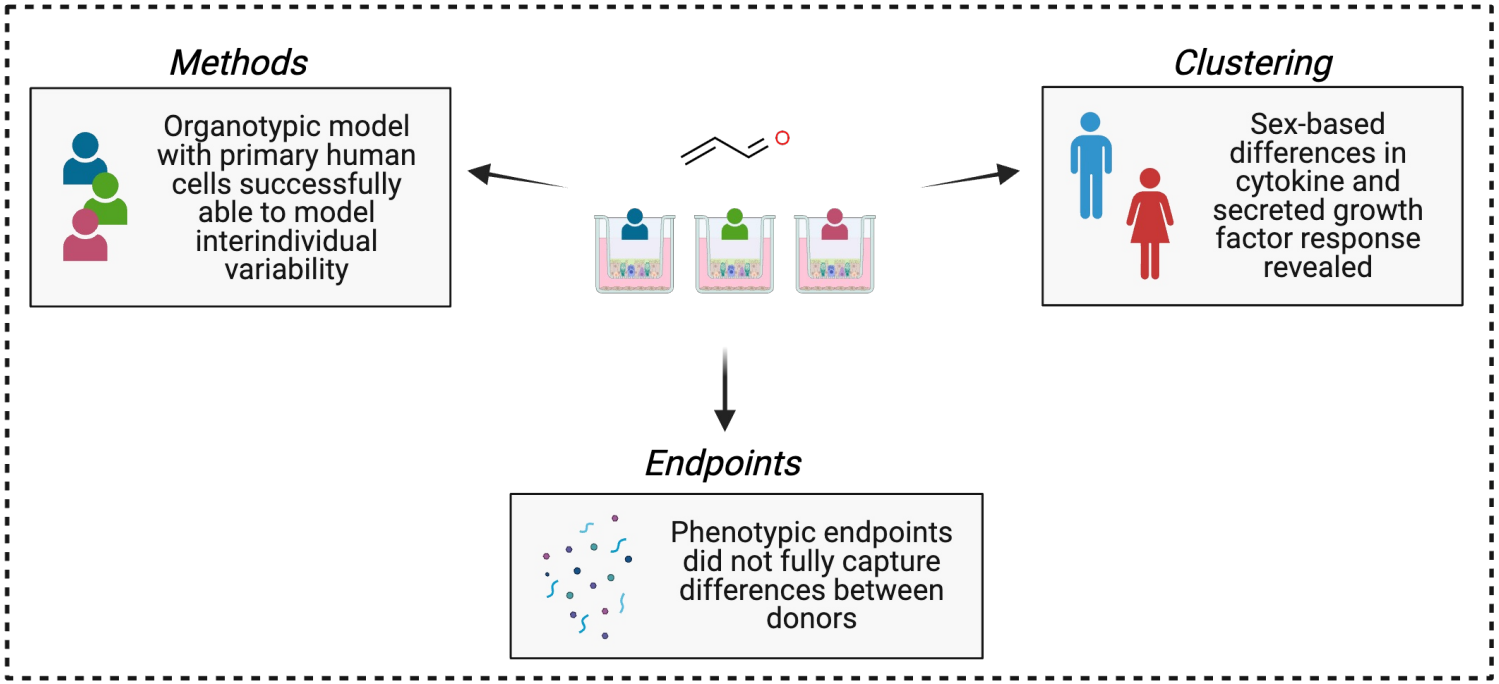


Input: Power curve model fit parameters for cytokine and growth factor data

	Cluster 1 (N=3)	Cluster 2 (N=5)	Cluster 3 (N=6)	P-value
Sex				
Female	0 (0%)	1 (20.0%)	4 (66.7%)	0.115
Male	3 (100%)	4 (80.0%)	2 (33.3%)	
Age				
Mean (SD)	26.0 (15.7)	41.0 (17.6)	28.4 (33.3)	0.652
Median [Min, Max]	19.0 [15.0, 44.0]	46.0 [13.0, 58.0]	13.5 [0.330, 91.0]	
BMD (Model Avg)				
Mean (SD)	2.90 (0.404)	4.85 (2.74)	3.08 (0.637)	0.203
Median [Min, Max]	2.79 [2.56, 3.35]	3.97 [3.21, 9.69]	3.05 [2.27, 3.86]	
BMD (Power Model)				
Mean (SD)	2.93 (0.434)	4.87 (2.73)	3.09 (0.640)	0.202
Median [Min, Max]	2.79 [2.57, 3.41]	4.00 [3.22, 9.70]	3.07 [2.28, 3.88]	

Manuscript in preparation

Conclusions



This study is impactful because it is among the first to combine in vitro primary co-culture models with advanced computational modeling to expand human response variability assessments in new approach methods (NAMs)-based risk assessment.

We detected factors underlying interindividual variability using machine learning.

Manuscript in preparation

Outline of Presentation

1. Share examples of recent efforts leveraging supervised and unsupervised machine learning to understand key biological mechanisms of inhaled toxicants in human clinical studies.

2. Highlight a study leveraging an organotypic *in vitro* co-culture model of the respiratory system to understand variables underlying interindividual variability in response to acrolein.

3. Discuss major takeaways, upcoming data science training efforts, and future studies.

Overarching Conclusions

Themes across all projects:

- Human respiratory toxicology data
- High interindividual variability
- Relatively small N and number of endpoints
- Goal of quantifying endpoints as a whole

Supervised and unsupervised machine learning represent methods that can aid in understanding key biological mechanisms of inhaled toxicants and interindividual variability in response to inhaled toxicant exposure.

Ongoing challenges:

- Sample size
- Human variability
- Batch effects
- Covariates
- Data pre-processing
- Selection and interpretation of ML
- Biases in analysis
- **Data analysis training**

Training the Next Generation of Toxicologists

- **inTelligence And Machine lEarning (TAME)** Toolkit, promoting didactic data generation, management, and analysis methods to **"TAME" data** in environmental health studies
- Development led by Dr. Julia Rager

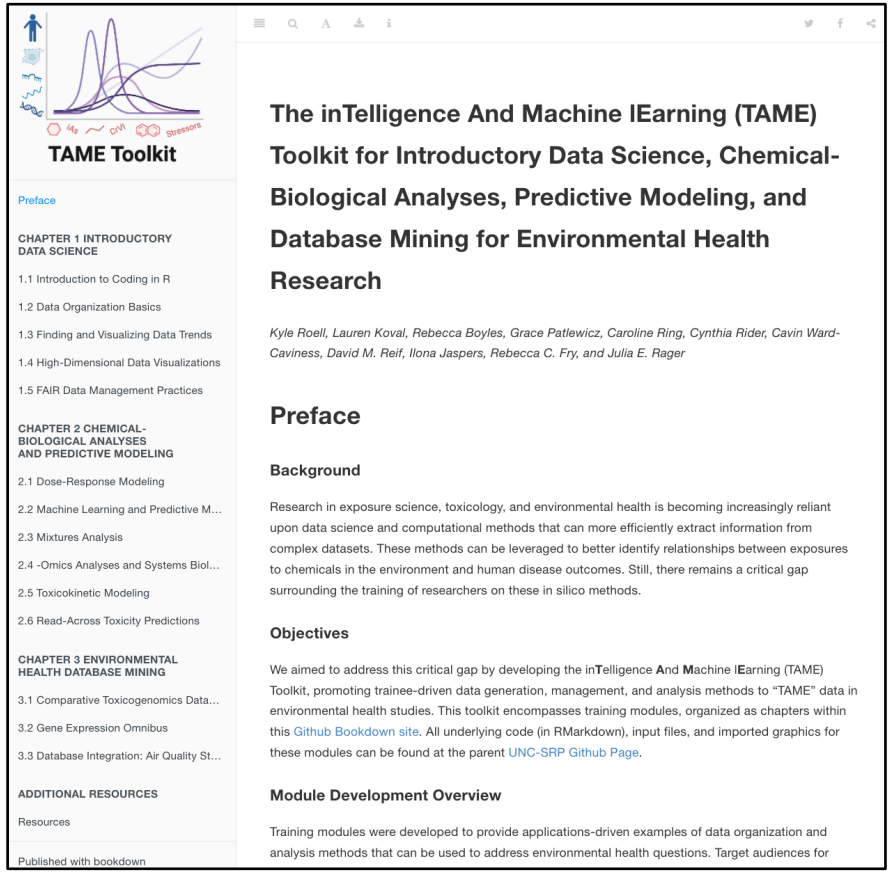
TECHNOLOGY AND CODE article

Front. Toxicol., 22 June 2022
Sec. Computational Toxicology and Informatics
Volume 4 - 2022 |
https://doi.org/10.3389/ftox.2022.893924

This article is part of the Research Topic
Computational Toxicology: Data Pipelines and Analysis
View all 4 Articles >

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*]



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...
2.5 Toxicokinetic Modeling
2.6 Read-Across Toxicity Predictions

CHAPTER 3 ENVIRONMENTAL HEALTH DATABASE MINING
3.1 Comparative Toxicogenomics Data...
3.2 Gene Expression Omnibus
3.3 Database Integration: Air Quality St...

ADDITIONAL RESOURCES
Resources

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.

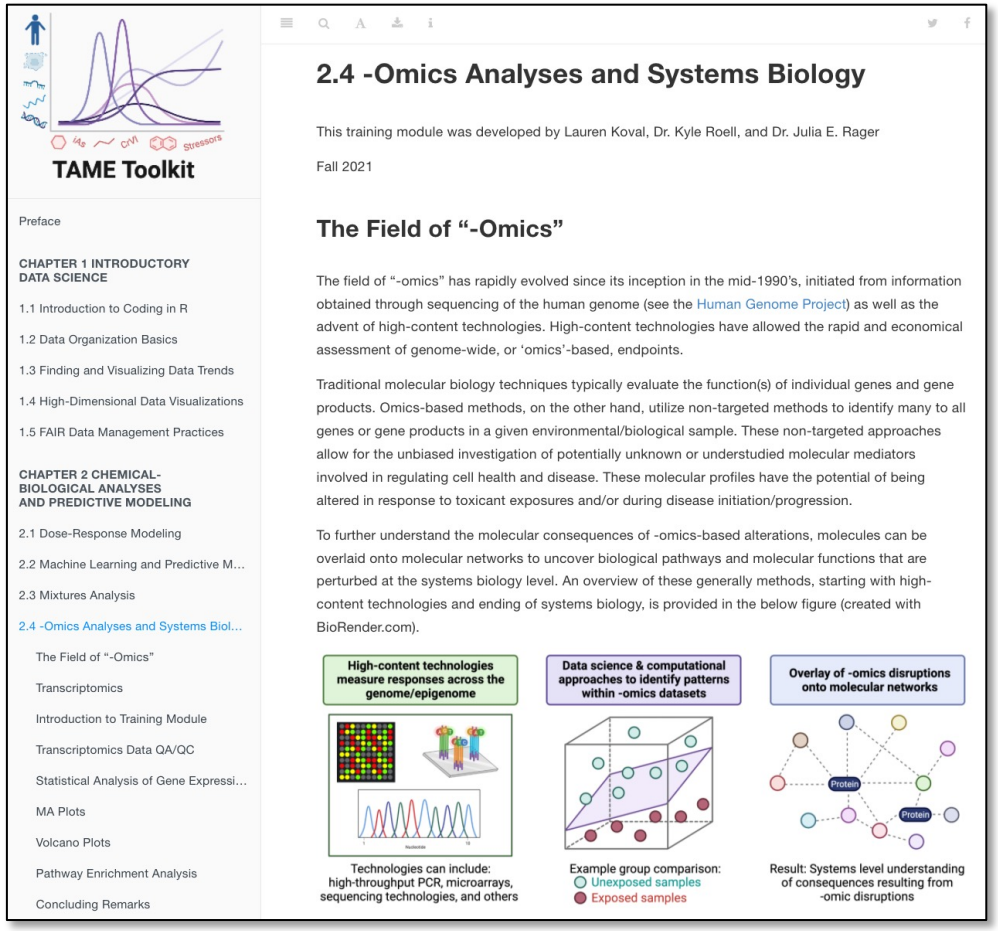
Module Development Overview

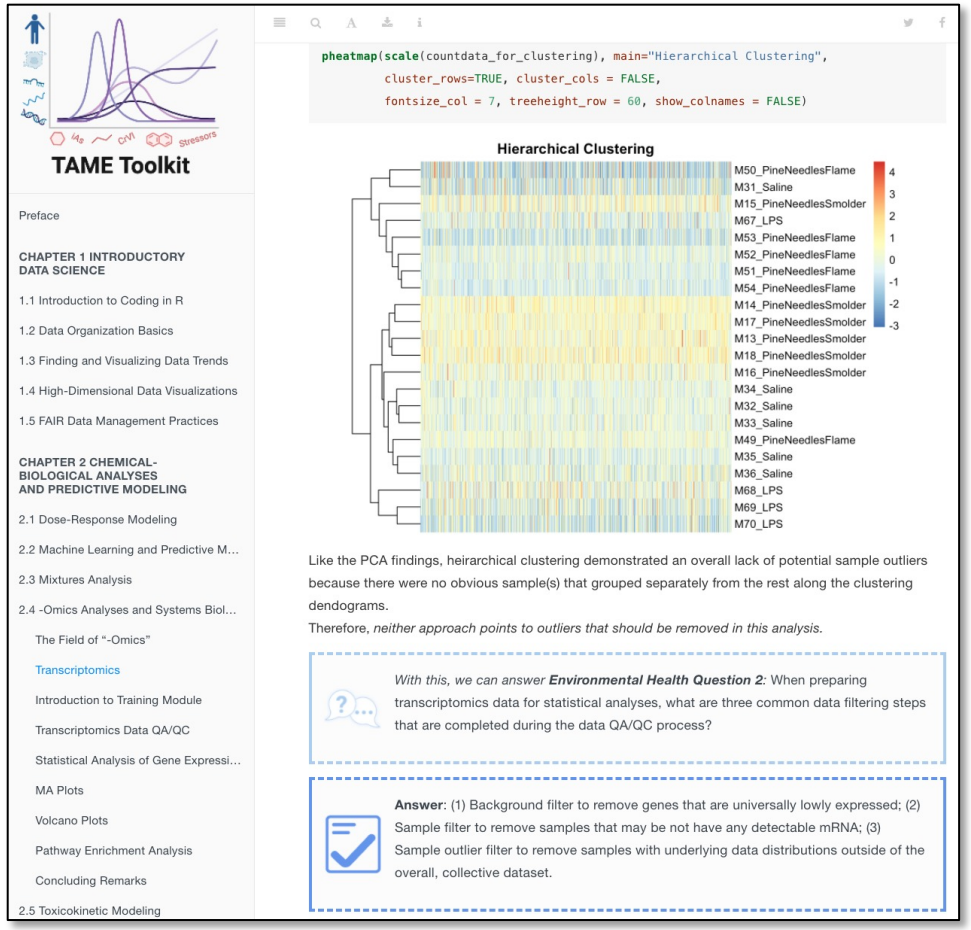
Training modules were developed to provide applications-driven examples of data organization and analysis methods that can be used to address environmental health questions. Target audiences for

Scan to be directed to TAME Site:

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

TAME is a Publicly Available, Online Bookdown Site

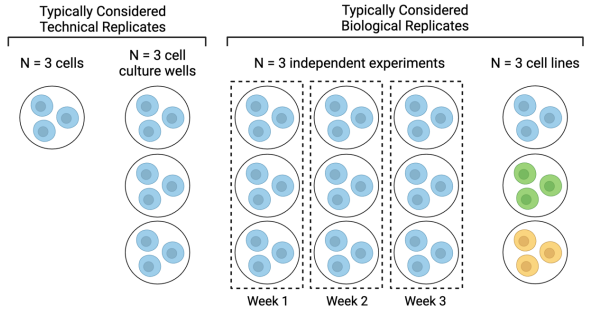


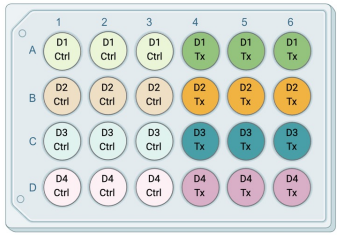


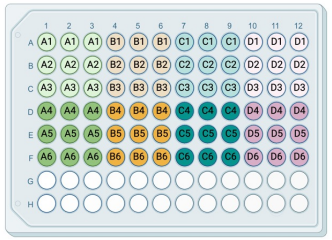
TAME 2.0 Coming Soon!

TAME 2.0 Chapter 4:
Converting Wet Lab Data Into Dry Lab Analyses

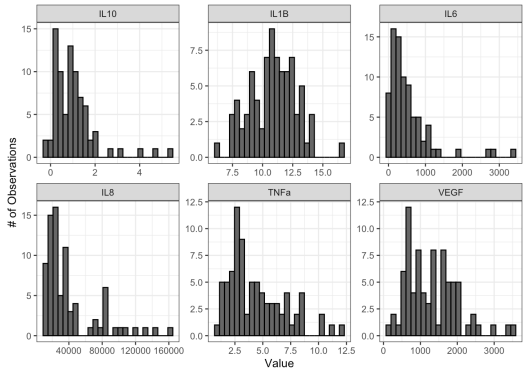
Experimental Design

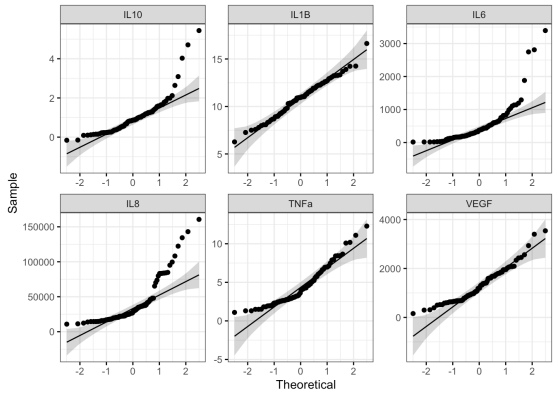




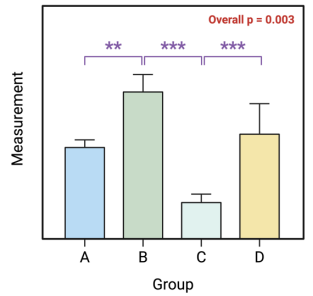


Data Processing & Transformation



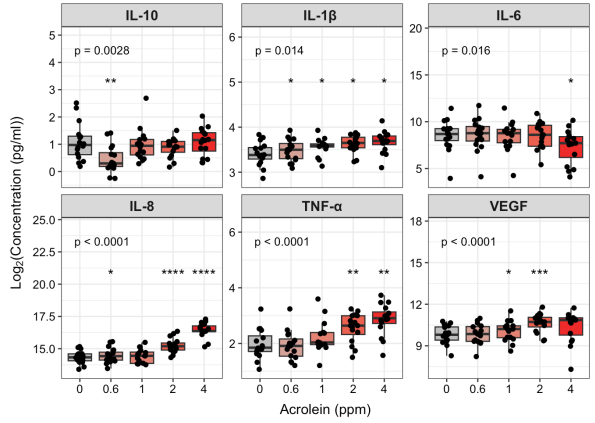


Basic Statistical Testing & Improved Visualizations



The **overall p-value** comes from the main statistical test (e.g., t-test, Wilcoxon test, ANOVA, Kruskal-Wallis, Friedman Test).

Pairwise p-values are derived from post-hoc tests such as pairwise t-tests, pairwise Wilcox tests, Tukey's HSD, and Dunn's test.

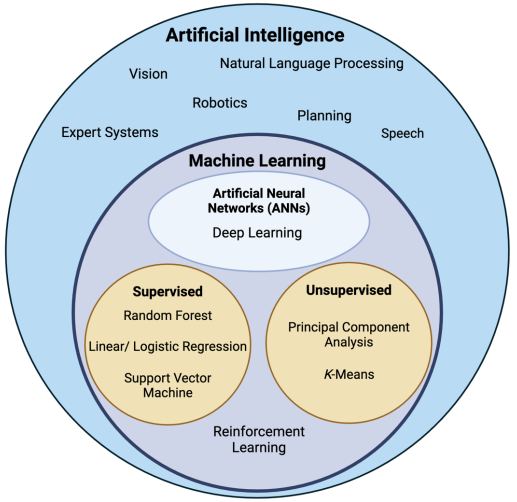


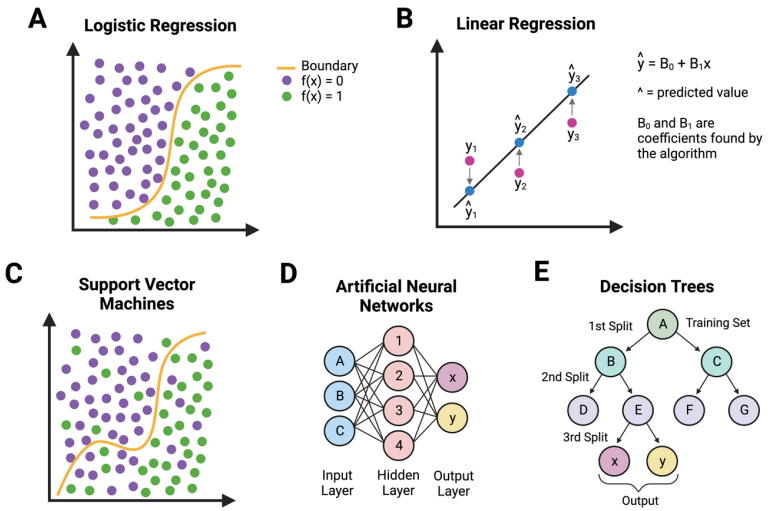
TAME 2.0 Chapter 5, Module 1:
Introduction to Machine Learning and Artificial Intelligence

Dr. David Reif

Alexis Payton

- General historical context and taxonomy of modern AI/ML, including ChatGPT!
- Application of machine learning in environmental health science
  - Why do we need machine learning?
  - Machine learning vs. traditional statistical methods
  - Predictive modeling in the context of environmental health science
  - Additional applications of machine learning in environmental health science
- Scripted examples of supervised and unsupervised machine learning in the following modules



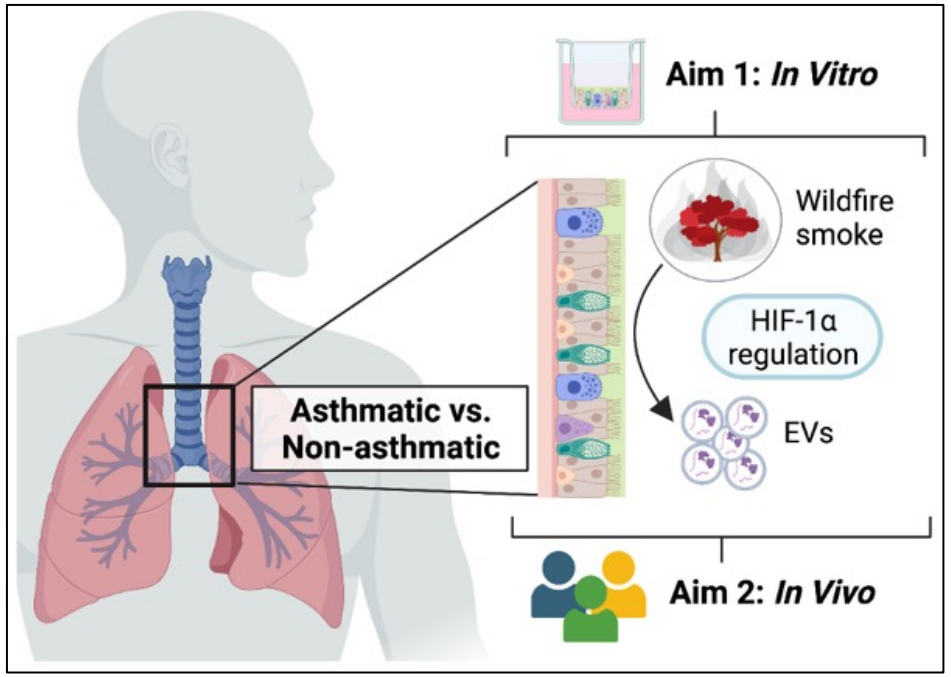


Upcoming Research

“Mechanisms of wildfire smoke toxicity and susceptibility involving extracellular vesicles in humans”

Goal: Determine differential responses to wildfire smoke exposure in asthmatics and non-asthmatics through the novel integration of EV signatures obtained from epithelial *in vitro* studies with clinical human *in vivo* studies on biomass smoke exposures.

We hypothesize that the hypoxia inducible factor 1 subunit alpha (HIF-1α) pathway mediates differential inflammatory responsiveness to biomass smoke exposure between asthmatics vs non-asthmatics through extracellular vesicle (EV)-mediated communication.



Acknowledgements

Rager Lab
Julia Rager, PhD
Alexis Payton

Alexis Lab
Neil Alexis, PhD
Heather Wells

Jaspers Lab
Ilona Jaspers, PhD
Parker Duffney, PhD (now EPA)
Stephanie Brocke
Aleah Bailey

CEMALB Study Coordinators
Carole Robinette
Martha Almond
Noelle Knight
Brian Ring

Study Participants

Other Collaborators
Agatha Ceppe, PhD (UNC)
Meghan Rebuli, PhD (UNC)
David Reif, PhD (NIEHS)
Shaun McCullough, PhD (RTI International)
Alysha Simmons, PhD (UNC)

Groups
SPIROMICS
TAME Contributors

Funding
CiTEM Training Grant (T32 ES007126)

National Heart, Lung, and Blood Institute
(R01 HL139369, P50 HL120100, F31 HL154758,
P50 HL120100)

TAME:

Thank you! Questions?

Contact: ehickman@email.unc.edu