United States Environmental Protection Agency

Predicting Chromatography-tandem Mass Spectrometry Amenability to Improve Non-targeted Analysis

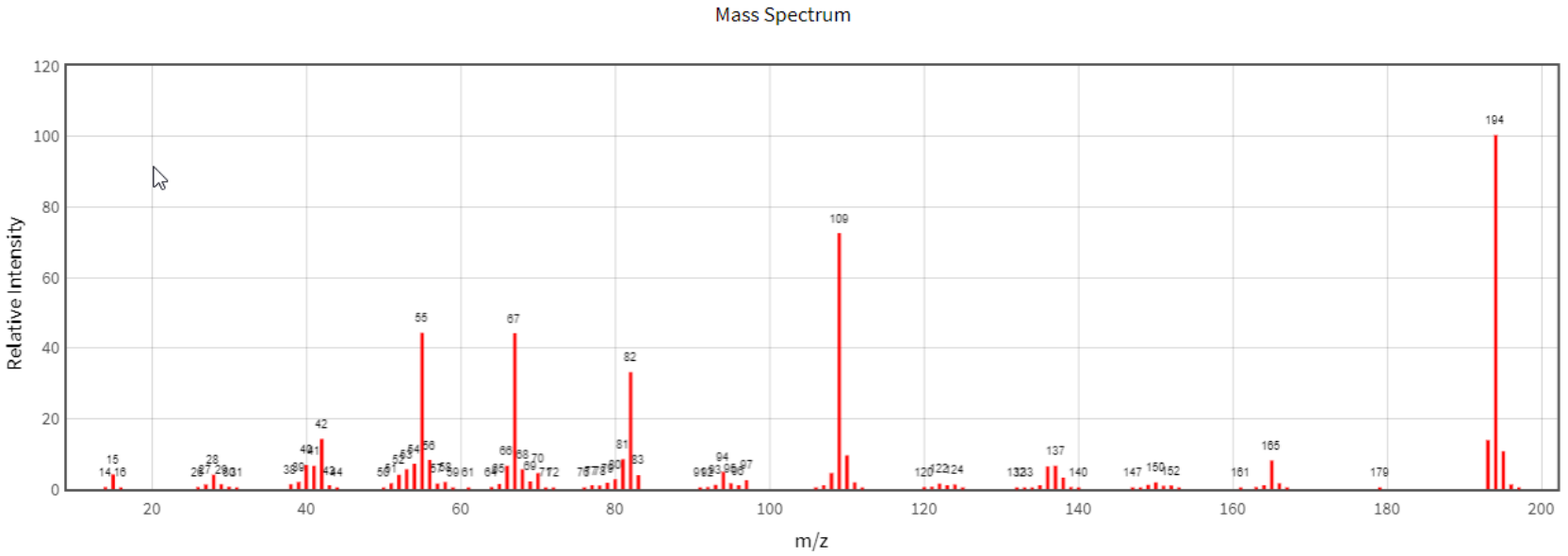
Charles Lowe[1], Kristin Isaacs[1], Chris Grulke[1], Jon Sobus[1], Elin Ulrich[1], Alex Chao[1,2], and Antony J. Williams[1]

1. Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, Research Triangle Park, NC
2. Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC

United States Environmental Protection Agency

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. This presentation has not been reviewed for policy and is not for distribution.

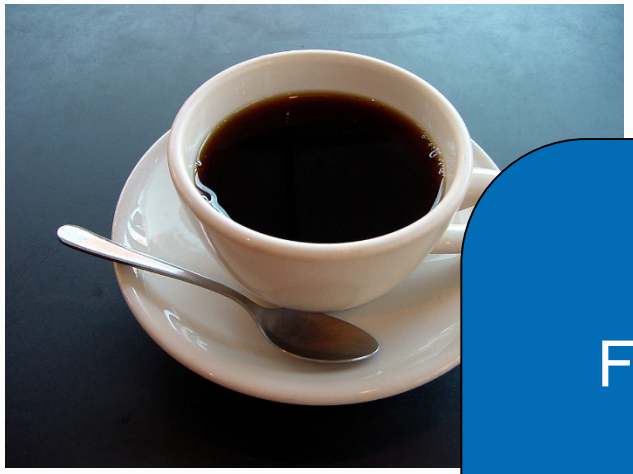
What are we trying to model?

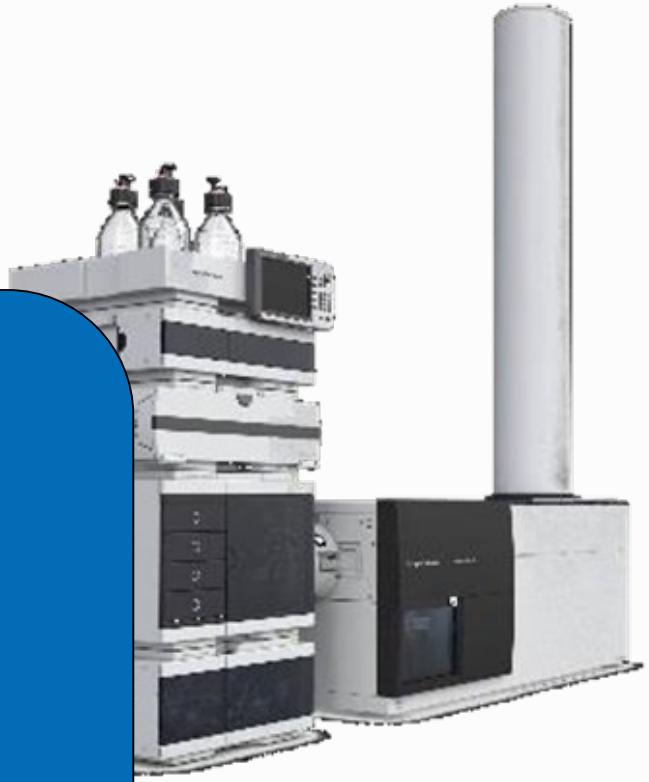


???

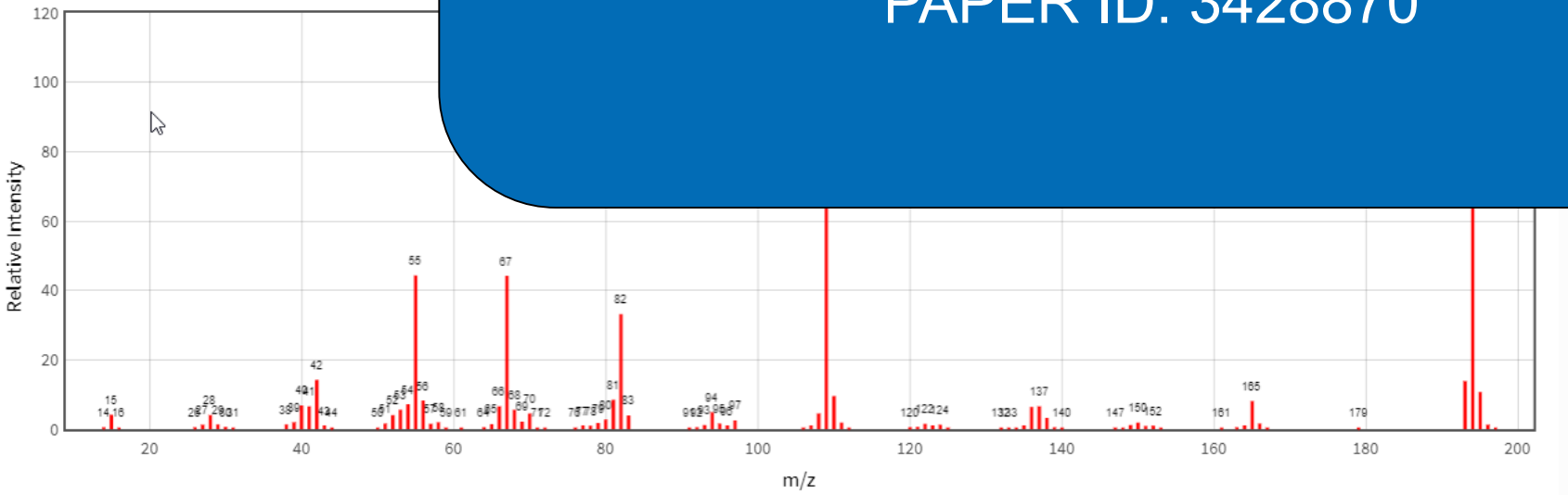
https://images.app.goo.gl/ftRmhwxEtZs95uKv7

What are we trying to model?





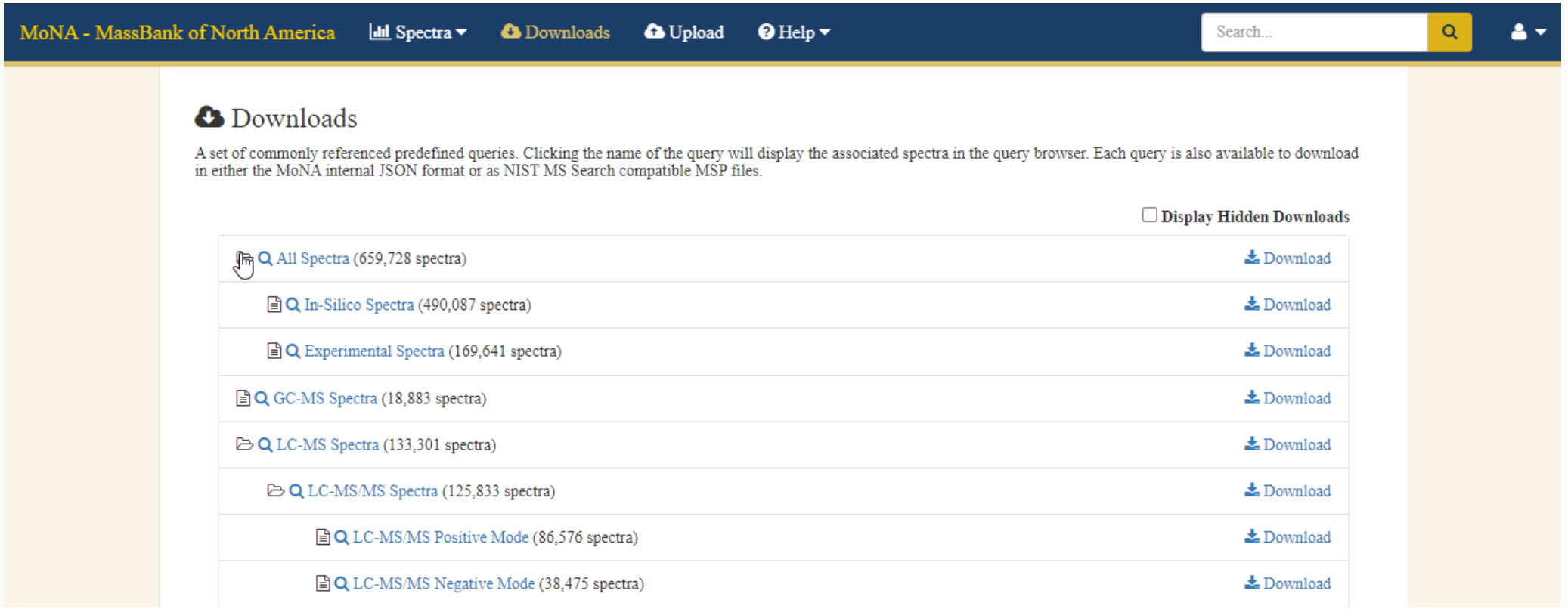
For more details on NTA at EPA, please see:
EPA's research initiatives on non-targeted analyses of environmental chemicals
PRESENTER: Jon Sobus
PAPER ID: 3428870



???

https://images.app.goo.gl/ftRmhwxEtZs95uKv7

The more data, the better (most of the time..)

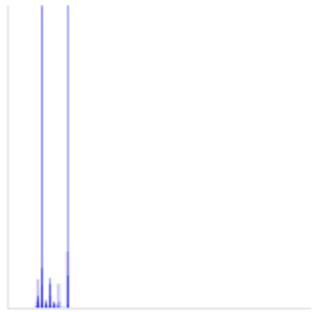


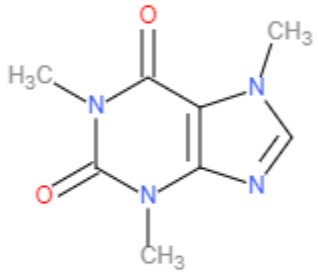
- 772 compounds in derivatized GCMS
- 7,199 compounds in non-derivatized GCMS
- **3,549 compounds in ESI+ LCMS**
- **2,630 compounds in ESI- LCMS**

EPA United States Environmental Protection Agency

Caffeine

Score: ★★★★☆



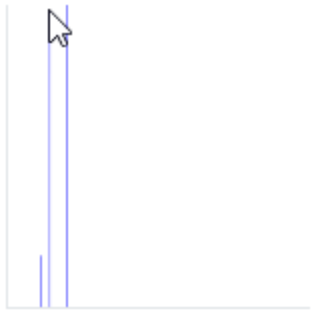


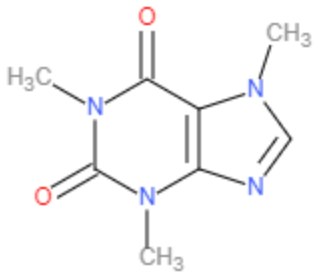
instrument	Pegasus III TOF-MS system...
instrument type	GC-EI-TOF
ms level	MS1
retention index	1880.2430
retention time	724.344 sec
ionization mode	positive
accession	OUF00133
date	2016.01.19 (Created 2010....
author	Tsujimoto Y, Tsugawa H, B...
license	CC BY-SA

Originally submitted to the MassBank High Quality Mass Spectral Database

Caffeine

Score: ★★★★☆

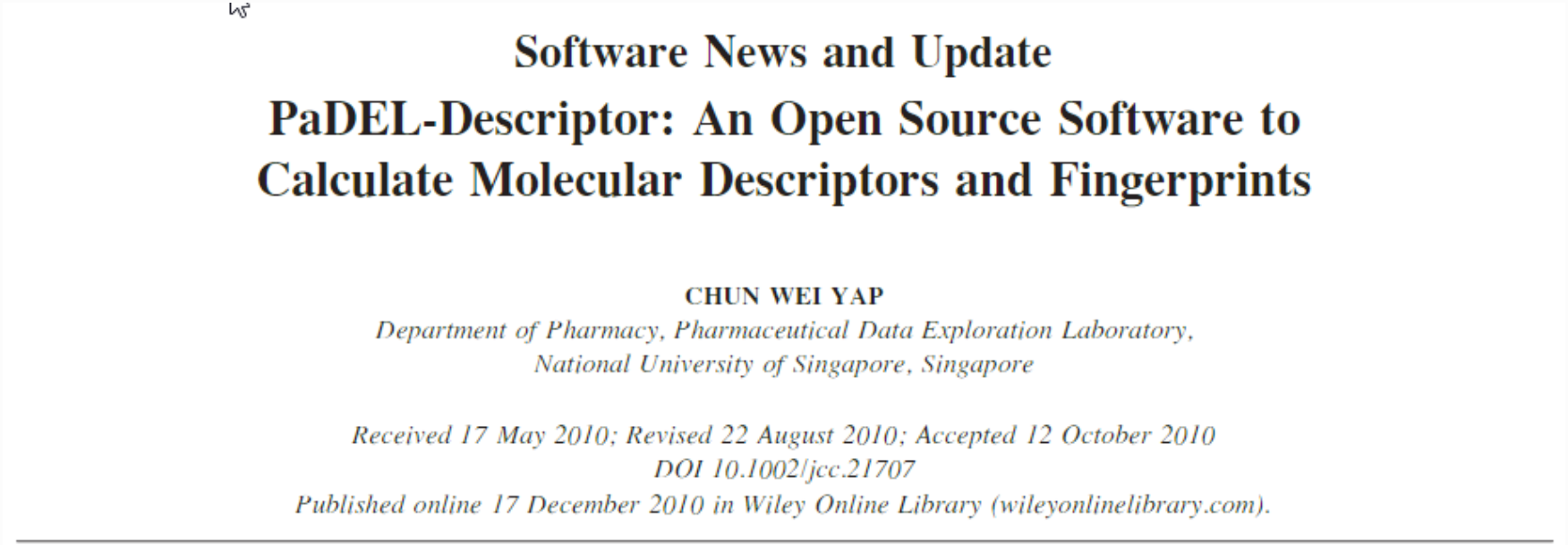




instrument type	QqQ
instrument	Micromass Quattromicro
collision energy	15eV
ionization	ESI
ionization mode	positive
ms level	MS2
precursor m/z	194.9000
precursor type	[M+H]+
accession	PM018511
publication	Alonso-Salces RM, Guillou...

Originally submitted to the RIKEN MS^n Spectral Database for Phytochemicals

Describing structures for modeling



Software News and Update

PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints

CHUN WEI YAP
Department of Pharmacy, Pharmaceutical Data Exploration Laboratory, National University of Singapore, Singapore

Received 17 May 2010; Revised 22 August 2010; Accepted 12 October 2010
DOI 10.1002/jcc.21707
Published online 17 December 2010 in Wiley Online Library (wileyonlinelibrary.com).

- 1,444 1D & 2D Molecular descriptors from QSAR-ready SMILES. Examples include…
  - Electrotopological state
  - McGowan volume (van der Waals volume)
  - molecular linear free energy relationships
  - Atom, bond, & ring counts
  - LogP predictions, etc..

Reduction of descriptor space

Dimension reduction will improve our models and make calculations quicker

1. Remove any constant descriptors (variance(x) = 0)
2. Remove nearly constant descriptors (SD < 0.25)
   - 0.25 gives a good balance between reduction and retention
3. Calculate pair-wise correlations between remaining descriptors
   - Eliminate based on a cutoff = 0.96 correlation

1,444 descriptors → 385 descriptors

Datasets suitable for modeling

Models need both training and test data

- 75% of data for training, 25% for testing
  - Data stratified to maintain proportions in outcome variable
  - Different for each model
  - InChIKey skeleton as identifier
- External validation datasets
  - EPA's NTA Collaborative Trial (ENTACT) data (explicitly removed from train/test sets)

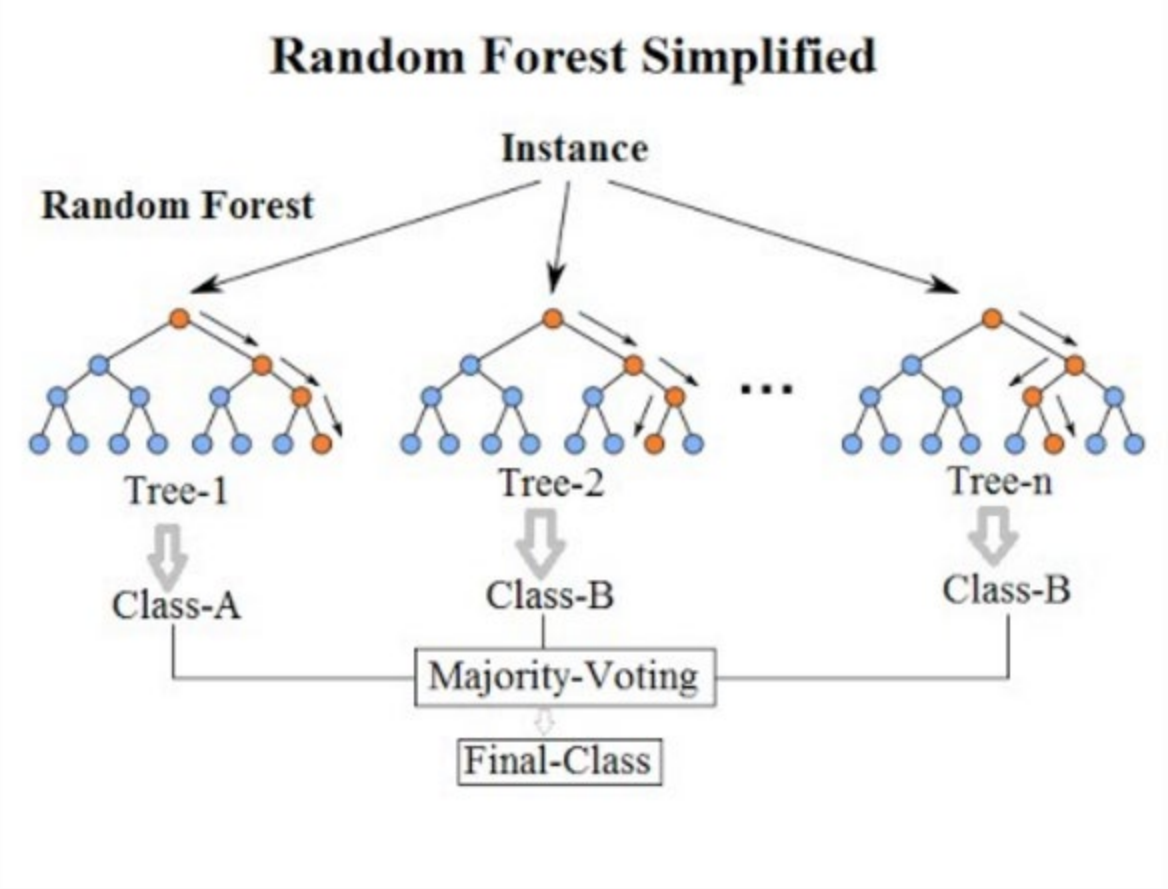
```
library(readxl)
library(caret)
library(randomForest)
library(funModeling)
library(tidyverse)
library(GA)
library(AdaSampling)
library(wsrf)
library(rsample)
library(dbscan)
```

R libraries used in study

Learning approach

- Four models
  - GC (derivatized), GC (not derivatized)

    ESI+ LC, **ESI- LC**
- Random forest (will explain)
  - Downsample absence data to match count of presence data
  - Optimize mtry and ntree via grid search
  - 5-fold cross validation
  - Y-randomization



Choosing the correct descriptors to predict the endpoint

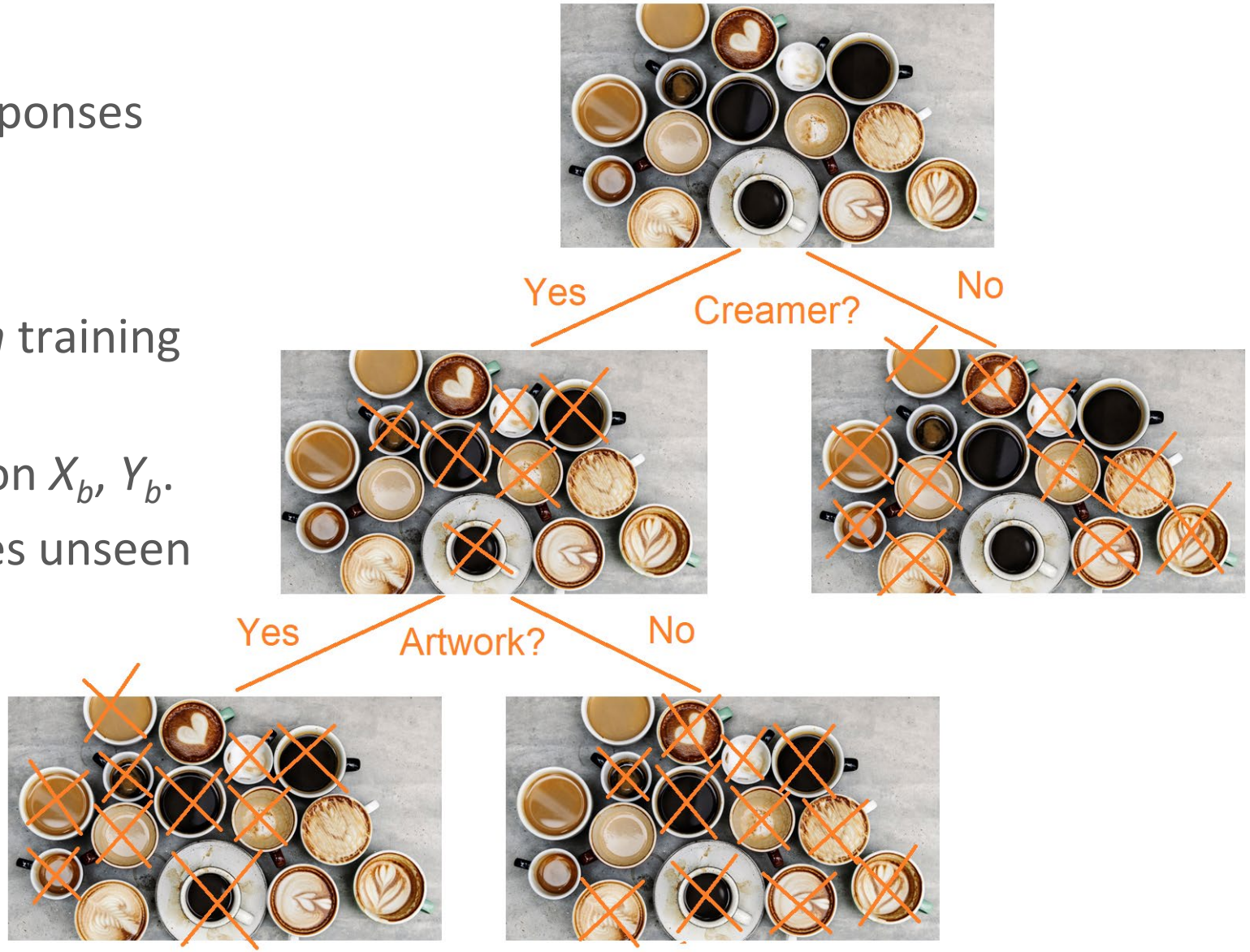
Random Forest Algorithm

Training set $X = x_1x_2...x_n$ with responses

$Y = y_1y_2...y_n$

For $b = 1,...,B$

1. Sample, with replacement, n training examples from X, Y; X_b, Y_b.
2. Train a classification tree f_b on X_b, Y_b.
3. The majority of all f_b classifies unseen samples.



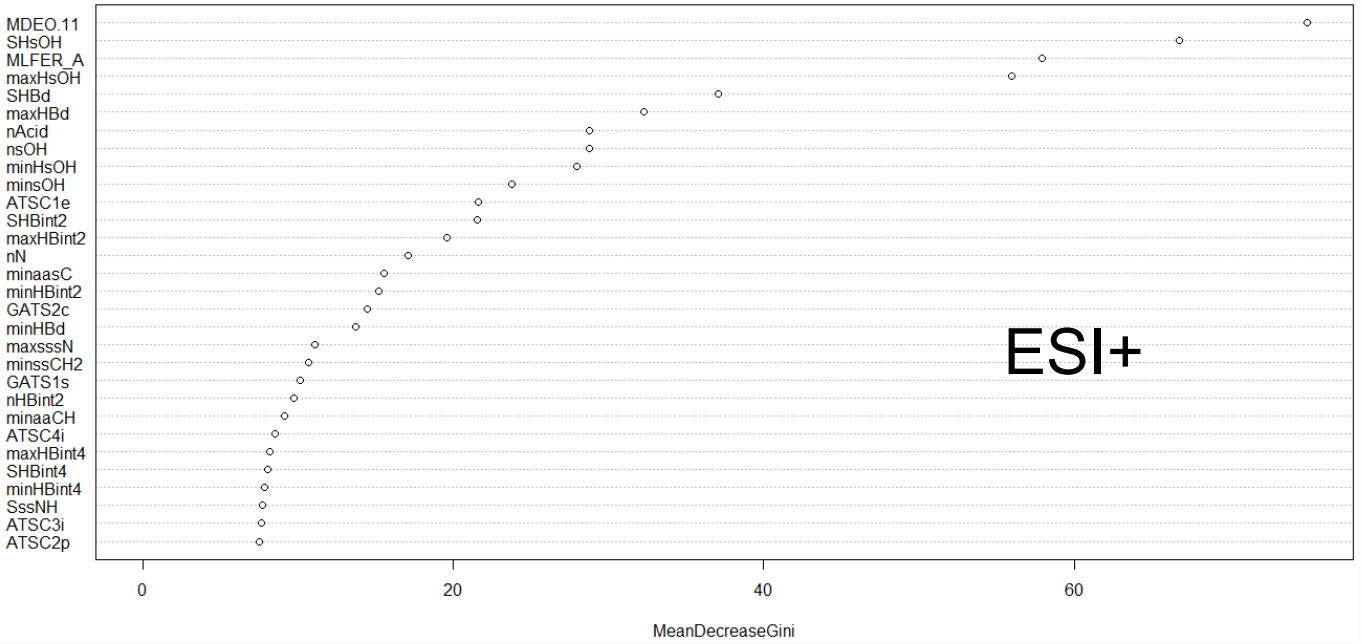
Negative data

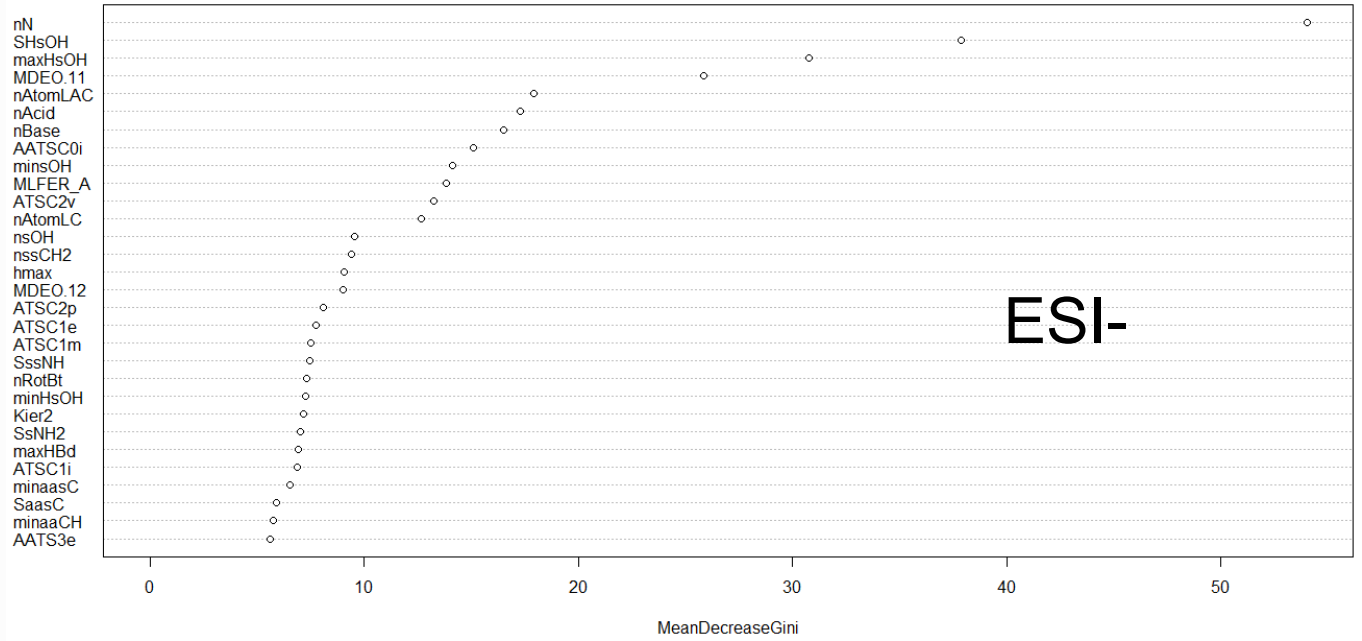
- Classification models need negative data, in addition to positive data
  - labs do not report chemicals NOT seen, only those identified by the instrument
- How do we provide a model with negative data?
  - produce the negative data ourselves (but note it is expensive)
  - assume all chemicals not present are absent
  - make assumption(s) as to what WAS tested
- For now, let's assume that if a chemical is detected in either ESI+/-, then it has also been tested in the other mode
- Still exploring reasonable assumptions for GCMS

Descriptor importance

Important descriptor descriptions

- MDEO-11
  - molecular distance edge between all primary oxygens
- MLFER-A
  - overall or summation solute hydrogen bond acidity
- SHsOH & maxHsOH
  - electrotopological state with respect to –OH fragments
- nN
  - the number of N atoms…





Model results

ESI+ Not Downsampled		
	Reference	
Prediction	Present	Absent
Present	2409	252
Absent	291	716
Sensitivity	0.8922	
Specificity	0.7397	
Balanced Accuracy	0.8159	

ESI+ Downsampled		
	Reference	
Prediction	Present	Absent
Present	2273	388
Absent	171	836
Sensitivity	0.9300	
Specificity	0.6830	
Balanced Accuracy	0.8065	

Model results

ESI- Not Downsampled		
	Reference	
Prediction	Present	Absent
Present	1659	305
Absent	291	1413
Sensitivity	0.8508	
Specificity	0.8225	
Balanced Accuracy	0.8366	

ESI- Downsampled		
	Reference	
Prediction	Present	Absent
Present	1649	315
Absent	271	1433
Sensitivity	0.8589	
Specificity	0.8198	
Balanced Accuracy	0.8393	

Internal test set results

ESI+ Not Downsampled		
	Reference	
Prediction	Present	Absent
Present	804	114
Absent	84	220
Sensitivity	0.9054	
Specificity	0.6587	
Balanced Accuracy	0.782	

ESI+ Downsampled		
	Reference	
Prediction	Present	Absent
Present	767	65
Absent	121	269
Sensitivity	0.8637	
Specificity	0.8054	
Balanced Accuracy	0.8346	

Internal test set results

ESI- Not Downsampled		
	Reference	
Prediction	Present	Absent
Present	551	104
Absent	115	452
Sensitivity	0.8273	
Specificity	0.8129	
Balanced Accuracy	0.8201	

ESI- Downsampled		
	Reference	
Prediction	Present	Absent
Present	545	92
Absent	121	464
Sensitivity	0.8183	
Specificity	0.8345	
Balanced Accuracy	0.8264	

Current & future work

- Comparing model results to ENTACT results
  - comparing predictions against independent labs, consensus of labs
- Considering new metrics for model quality
  - balanced accuracy not ideal when negative data may contain false negatives
- Applicability domains for models under development
  - global and local measures
- Working with collaborators to improve available data
  - data from additional potential collaborators would be GREATLY appreciated

Contributing researchers

EPA ORD

Hussein Al-Ghoul*
Alex Chao*
Louis Groff*
Jarod Grossman*
Chris Grulke
Kristin Isaacs
Sarah Laughlin*
Jon Sobus
Kamel Mansouri*
James McCord
Andrew McEachran*
Jeff Minucci
Seth Newton
Katherine Phillips

EPA ORD (cont.)

Tom Purucker
Ann Richard
Randolph Singh*
Mark Strynar
Elin Ulrich
John Wambaugh
Antony Williams

GDIT

Ilya Balabin
Tom Transue
Tommy Cathey

* = ORISE/ORAU


Thank you for
Listening!

Questions?

Email: lowe.charles@epa.gov