TAP1 and TAP2 Preliminary Results:

Brief summary of LC-MS^1 and LC-MS^2 data processing/analysis, and multivariate statistical analysis of LC-MS^1 data

Seth R. Newton and John T. Sloop

Thursday, March 16, 2022

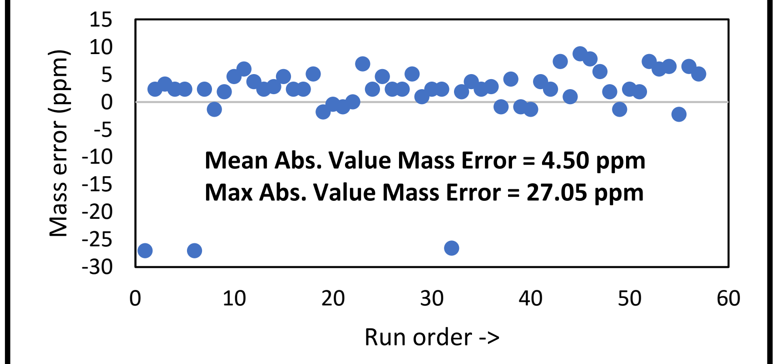
Goals of this project

- Use LC-MS and GC-MS instrumentation to gather as much data as possible on chemicals present in drinking water samples from homes across the state of California
  - Interested in link between chemicals present in drinking water and breast cancer
- From the data collected, use (a) toxicity values, abundance, and detection frequency and (b) results from multivariate stats modeling to define lists of “features of importance” for further validation of chemical identity
  - Either by *de novo* NTA or confirmation with standards via targeted analytical methods

Data processing workflow

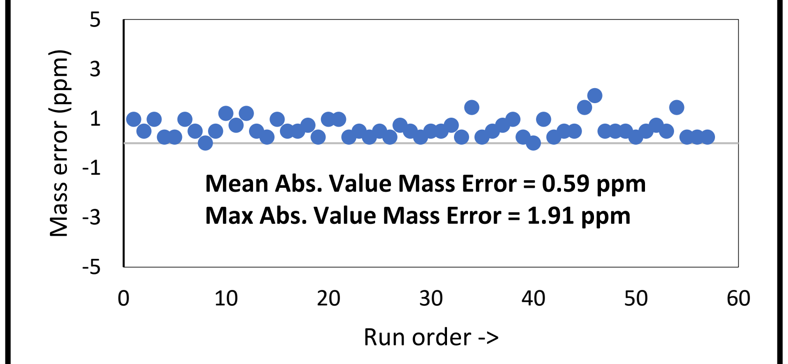
- LC-MS^1 Data:
  - Profinder for feature extraction, Mass Profiler Professional (MPP) for matching to MS-Ready formula, NTA WebApp to automate searching of features on dashboard
  - NTA WebApp results are then analyzed by categorizing features based on:
    - Number of Data Source Hits;
    - Availability of Toxicity Data;
    - And then (for those with Tox Data available) by ToxPi score
- LC-MS^2 Data:
  - Personal compound databases and libraries (PCDL) matching using Qualitative Analysis
- Results can be further confirmed by matching MS^1 results with MS^2 experimental data (and GC results)

QA/QC: Tracer mass error (ppm)



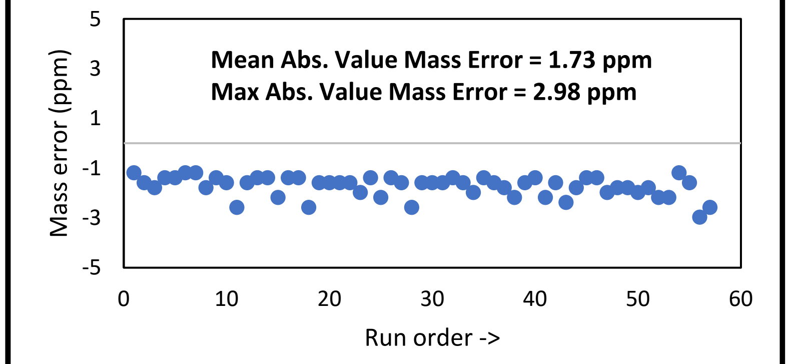
13C3 Atrazine (ESI+)

Mean Abs. Value Mass Error = 4.50 ppm
Max Abs. Value Mass Error = 27.05 ppm



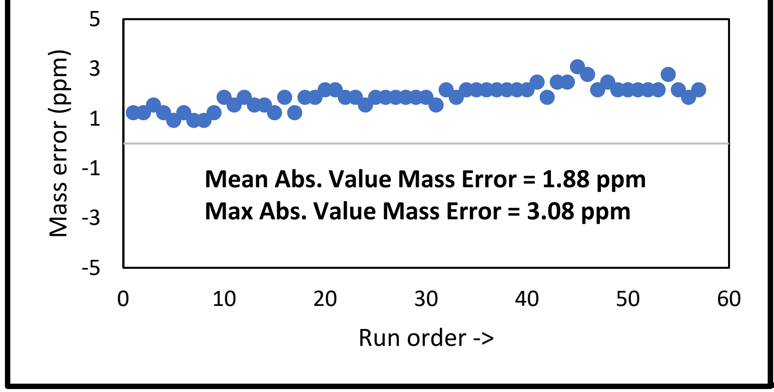
13C4 MPFOA (ESI-)

Mean Abs. Value Mass Error = 0.59 ppm
Max Abs. Value Mass Error = 1.91 ppm



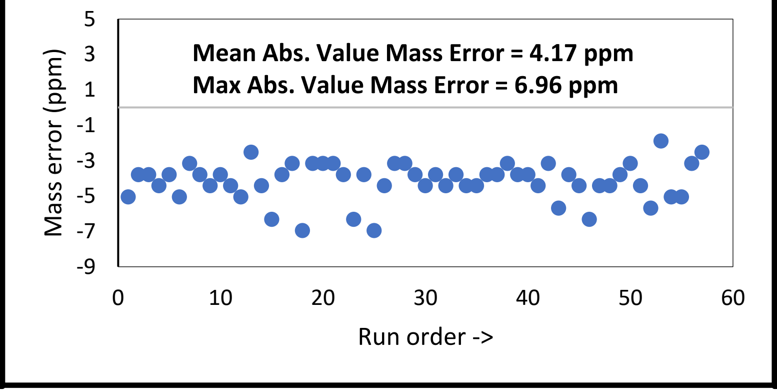
13C4 MPFOS (ESI-)

Mean Abs. Value Mass Error = 1.73 ppm
Max Abs. Value Mass Error = 2.98 ppm



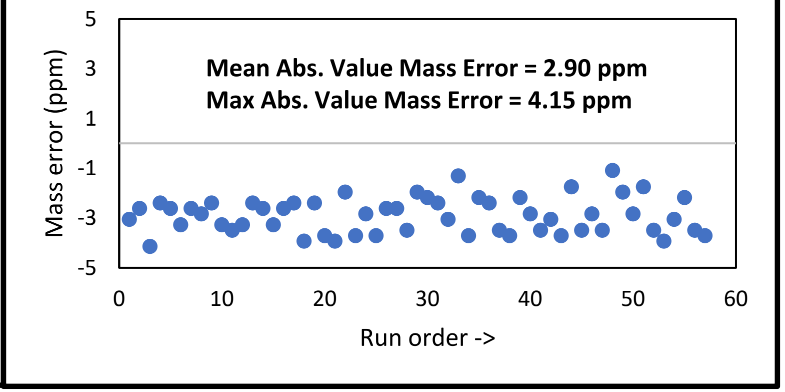
D3 Pyriproxyfen (ESI+)

Mean Abs. Value Mass Error = 1.88 ppm
Max Abs. Value Mass Error = 3.08 ppm



13C6 Methyl Paraben (ESI-)

Mean Abs. Value Mass Error = 4.17 ppm
Max Abs. Value Mass Error = 6.96 ppm

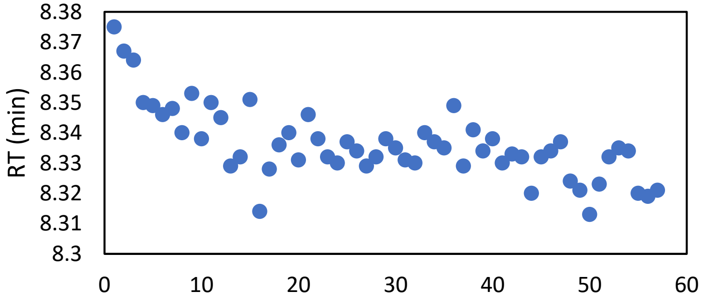


13C4 15N2 Fipronil sulfone (ESI-)

Mean Abs. Value Mass Error = 2.90 ppm
Max Abs. Value Mass Error = 4.15 ppm

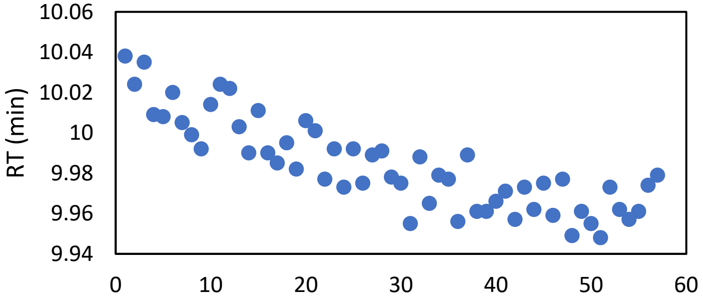
QA/QC: Tracer retention times (RT, min)

13C3 Atrazine (ESI+)



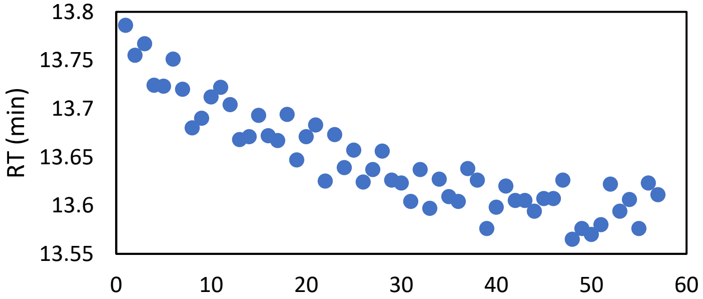
Avg. RT = 8.34 min
Range = 0.06 min

13C4 MPFOA (ESI-)



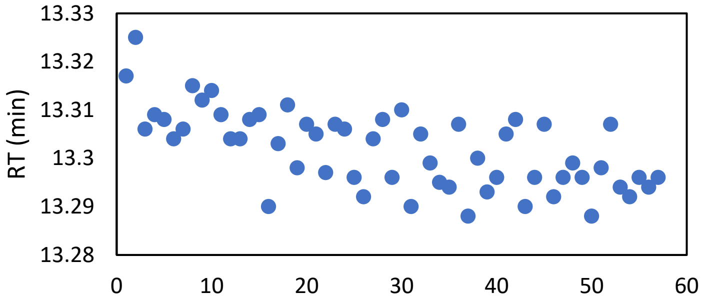
Avg. RT = 9.98 min
Range = 0.09 min

13C4 MPFOS (ESI-)



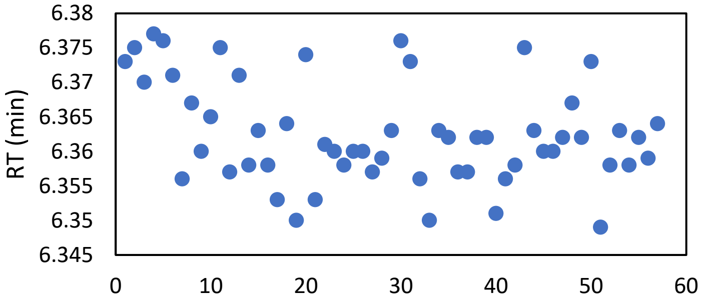
Avg. RT = 13.65 min
Range = 0.22 min

D3 Pyriproxyfen (ESI+)



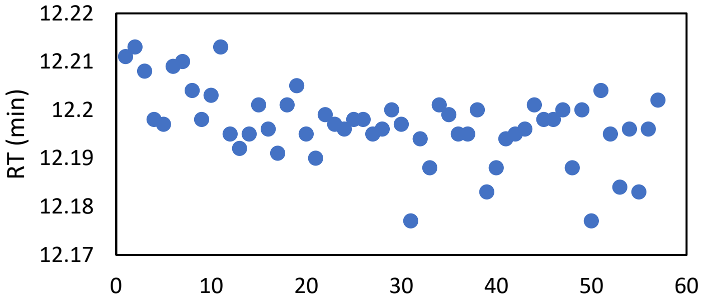
Avg. RT = 13.30 min
Range = 0.04 min

13C6 Methyl Paraben (ESI-)



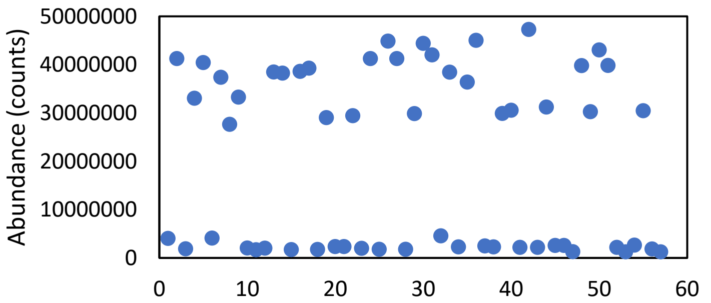
Avg. RT = 6.36 min
Range = 0.03 min

13C4 15N2 Fipronil sulfone (ESI-)



Avg. RT = 12.20 min
Range = 0.04 min

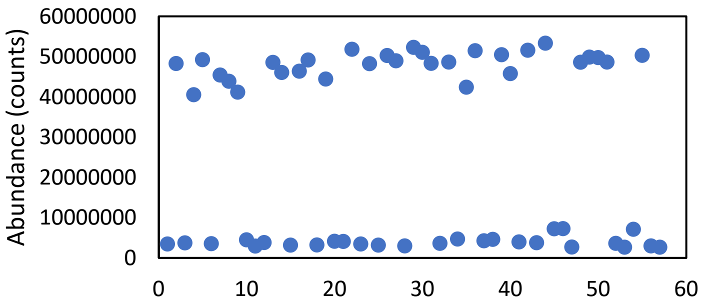
QA/QC: Tracer intensity (counts)



13C3 Atrazine (ESI+)

TAP1 Mean = 2,245,797
CV = 36 %

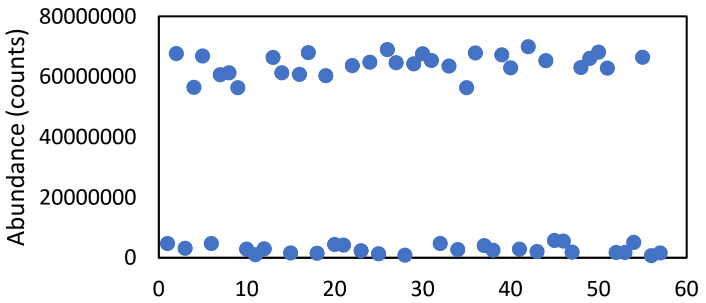
TAP2 Mean = 37,046,276
CV = 15 %



13C4 MPFOA (ESI-)

TAP1 Mean = 3,951,026
CV = 33 %

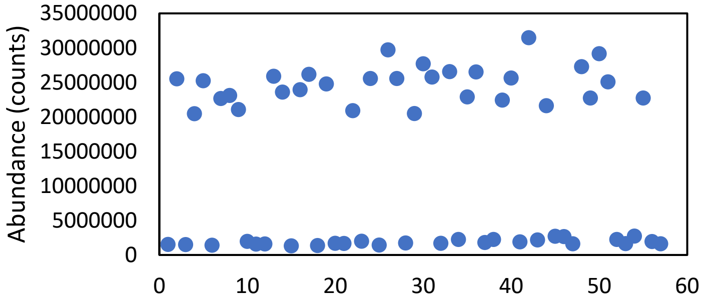
TAP2 Mean = 48,135,598
CV = 7 %



13C4 MPFOS (ESI-)

TAP1 Mean = 2,865,087
CV = 53 %

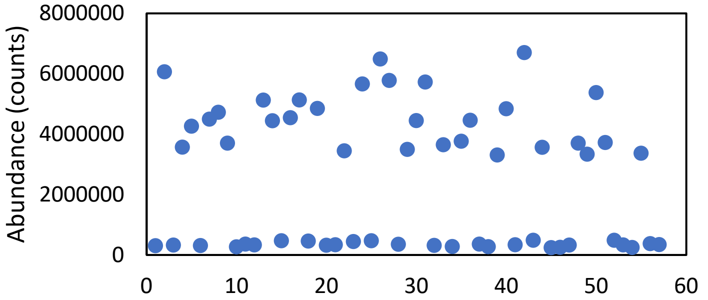
TAP2 Mean = 64,125,775
CV = 6 %



D3 Pyriproxyfen (ESI+)

TAP1 Mean = 1,822,698
CV = 22 %

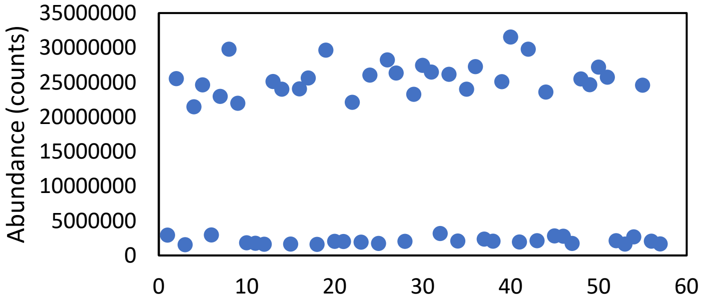
TAP2 Mean = 24,711,088
CV = 11 %



13C6 Methyl Paraben (ESI-)

TAP1 Mean = 343,128
CV = 22 %

TAP2 Mean = 4,519,686
CV = 22 %



13C4 15N2 Fipronil sulfone (ESI-)

TAP1 Mean = 2,091,057
CV = 23 %

TAP2 Mean = 25,645,028
CV = 10 %

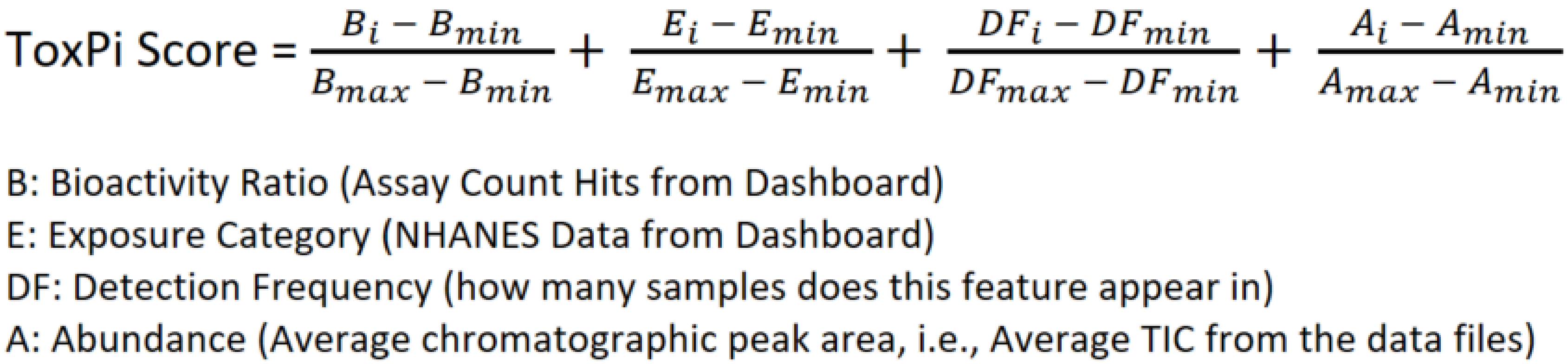
(Tracers incorrectly spiked at different concentrations for TAP1 and TAP2 samples)

NTA WebApp Initial Results

- Performed feature extraction in Profinder and MS-Ready formula matching in MPP
- Total of 948 formulas input into the WebApp
- After WebApp filtering and processing, left with 664 unique formulas with formula match score > 85
- Now, need to set priority of features to investigate
  - 15,049 total potential candidates

ToxPi Scoring

- Previous ToxPi calculation [1]:



$$\text{ToxPi Score} = \frac{B_i - B_{min}}{B_{max} - B_{min}} + \frac{E_i - E_{min}}{E_{max} - E_{min}} + \frac{DF_i - DF_{min}}{DF_{max} - DF_{min}} + \frac{A_i - A_{min}}{A_{max} - A_{min}}$$

B: Bioactivity Ratio (Assay Count Hits from Dashboard)
E: Exposure Category (NHANES Data from Dashboard)
DF: Detection Frequency (how many samples does this feature appear in)
A: Abundance (Average chromatographic peak area, i.e., Average TIC from the data files)

- Basically, boils down to a toxicity term, detection frequency term, and abundance term
- However, for this work, potentially rethink the individual terms used in this calculation

[1] Newton et al., Environmental Pollution, 234 (2018) 297-306.

ToxPi Scoring

- Searched every candidate's DTXSID on CompTox Dashboard via Batch Search
  - Pulling back TEST, Assay Hit Counts, Data Source Hits
- From total of 15,049 candidates, the following information/metadata is available:
  - TEST (DevTox, Ames, OralRat): 9,122 candidates
  - ToxCast: 294 candidates
- Lack of ToxCast data for most of the candidates, so considering multiple approaches for the ToxPi calculation
  - *Will determine which approach we take after all data collection is complete, considering the recommendations of collaborators*

Method 1 ToxPi Calculation (using TEST data)

- $ToxPi = 2\,T + 1.5\,A + 0.5\,DF$
- $T = \left(\frac{1}{3}\right) DevTox + \left(\frac{1}{3}\right) Ames + \left(\frac{1}{3}\right) OralRat$
- $A = \frac{Max_{Feature_i}}{Max_{AllFeatures}}$
- $DF = \frac{Y_i - Y_{min}}{Y_{max} - Y_{min}}$

All three of these values are based on mammalian studies, so all three are assigned the same weight.

Previously, average abundance for a feature across all samples was used to determine this value. In this study, we're more interested in the maximum measured abundance for a feature given any sample when assigning importance for further investigation.

Method 2 ToxPi Calculation (using ToxCast data)

- $ToxPi = 2\,T \;+\; 1.5\,A \;+\; 0.5\,DF$
- $T = \frac{X_i}{X_{max}}$
  - $X = AH_{ratio} \times \sqrt{AH}$
- $A = \frac{Max_{Feature_i}}{Max_{AllFeatures}}$
- $DF = \frac{Y_i - Y_{min}}{Y_{max} - Y_{min}}$

Imagine a scenario where there are two features being compared. Feature 1 was tested in 500 assays and found active in 50, and Feature 2 was tested in 10 assays and found active in 1. If just using the ratio, these two features are assigned the same value for their toxicity term, being 0.1. We think the raw number of active assays for Feature 1 should be taken into consideration when scoring and ranking these features.

Classification method of potential candidates

- Six possible sub-groups based on:
  - (i) availability of toxicity date (yes = A, no = B)
  - (ii) data source hits (top = 1, not = 2)
  - (iii) ToxPi score (top = α, not = β)
- A1α: Toxicity data available, largest Data Source hits, largest ToxPi score
- A1β: Toxicity data available, largest Data Source hits, not largest ToxPi score
- A2α: Toxicity data available, not largest Data Source hits, largest ToxPi score
- A2β: Toxicity data available, not largest Data Source hits, not largest ToxPi score
- B1: No toxicity data available, largest Data Source hits (no ToxPi score)
- B2: No toxicity data available, not largest Data Source hits (no ToxPi score)

Candidate grouping: Results

	TEST data		**ToxCast data**	
Classification	Hits	MS^2 matches	Hits	MS^2 matches
A1α	89	2	80	14
A1β	164	13	26	2
A2α	164	0	26	0
A2β	7,739	4	125	3
B1	290	9		
B2	5,024	1		

ToxPi Summary

- Found total of 89 A1α candidates based on TEST data, and 80 A1α candidates based on ToxCast data
  - “Most interesting” candidates based on highest ToxPi score
  - “Most likely” candidates based on highest Data Source Hits
- Need to:
  - Finish processing remaining samples and collecting data (once method development is *completely* finished)
  - Continue to finish processing and analyzing GC-MS results
- Eventually, will determine which ToxPi method we will use

Exploratory multivariate stats approach

- Using principal components analysis (PCA) as a tool for data reduction and visualization is routinely used [2,3]
  - PCA is unsupervised technique, meaning no information about sample "response" is used in the process (groupings/clusters based on response occur naturally, *i.e.,* unsupervised)
- Performed PCA and random forests (RF) classification to determine if samples will group together by geographic region, and if so, which variables (chemicals) are most responsible for this separation
  - Possible clustering/grouping ideas: based on geographic location, homes of individuals with/without breast cancer, drinking water provider, etc.

[2] B. Everitt, T. Hothorn, Springer, New York, NY, 1 (2011); [3] J.T. Sloop et al., JTEMB 54 (2019) 62-68.

Random forests (RF) modeling

- Supervised machine learning technique that can be used for classification and regression
- RF combines hundreds or thousands of decision trees
  - Each decision tree is trained on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features
  - The final predictions of the random forest are made by averaging the predictions of each individual tree

Random forest feature importance (RFI) value

- Random forest classification measures feature importance in two ways: permutation importance and *gini* importance [4,5]
- *Gini* index (GI)
  - Criterion used when growing data trees in random forest classification
  - The *gini* importance measures the significance of a feature in relation to a tree and a split in the random forest ensemble of trees
- The higher the value for *gVIj*, the better the feature was in splitting the data, and the greater the significance of that feature [6]

$$gVI_j = \frac{1}{ntree}\sum_{k=1}^{ntree} gVI_{jk}$$

[4] L. Breiman, Mach. Learn. 45 (2001) 5-32; [5] J. Carter et al., Expert Syst. Appl. 115 (2019) 245-255;
[6] B.A. Goldstein et al., Stat. Appl. Genet. Mol. Bio. 10 (2011) 1-36.

Principal components analysis (PCA)

- Unsupervised, multivariate technique with the main goal of reducing dimensionality
- First principal component is the linear combination of the original set of variables whose sample variance is greatest amongst all sets of linear combinations
  - $y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1q}x_q$
- 2-D and 3-D PCA are commonly performed
  - This visualization can be used to help determine if samples do form inherent groups (scores plot) and the chemicals of most importance when separating clusters of samples (loadings plot)

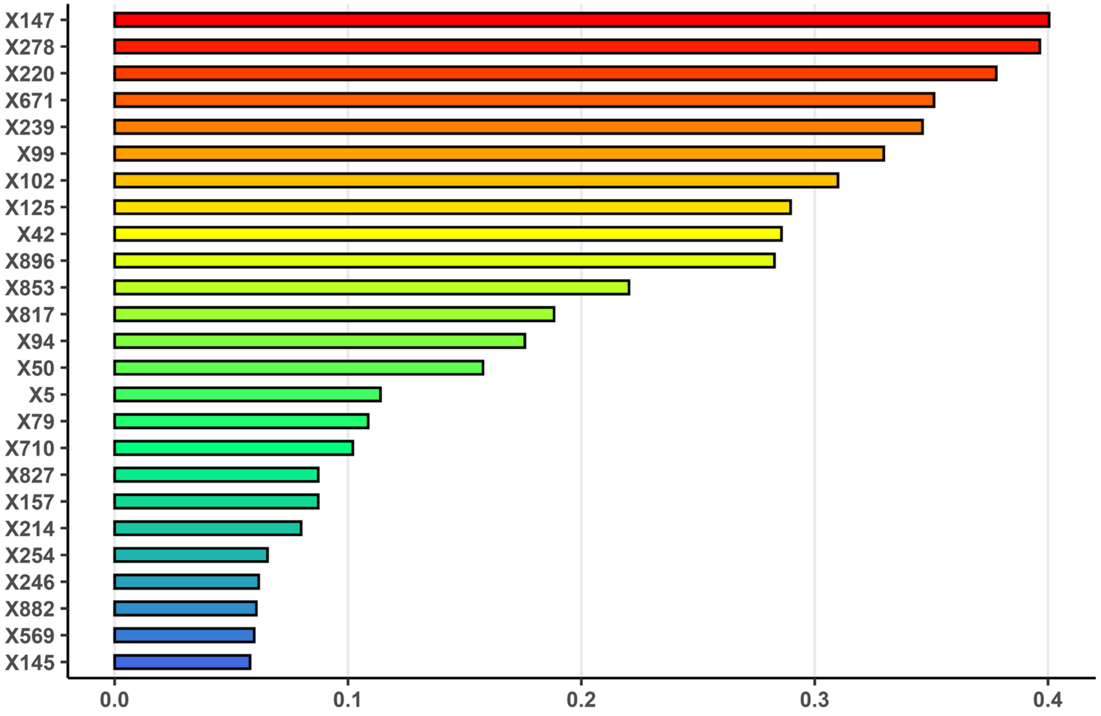
[1] B. Everitt, T. Hothorn, Springer, New York, NY, 1 (2011).

Data processing approaches

- All work shown here was performed on all extracted features and the A1α chemicals from using ToxPi Method 2 (ToxCast data)
  - 622 total features from set of all features
  - 80 features from A1α group of ToxPi Method 2
- Wanted to only use instrumental response for each chemical candidate and no other metadata or variables (Data Source hits, ToxPi, etc.)
  - No imputation, transformation, or standardization performed prior to RF (tree-based algorithm, not sensitive to scale)
  - No imputation or transformation, but center and scale to mean = 0, st.dev = 1 prior to PCA

Random Forests Feature Importance (RFI) Results

(Top 25 Features from A1α Data)



RF model performance using A1α features

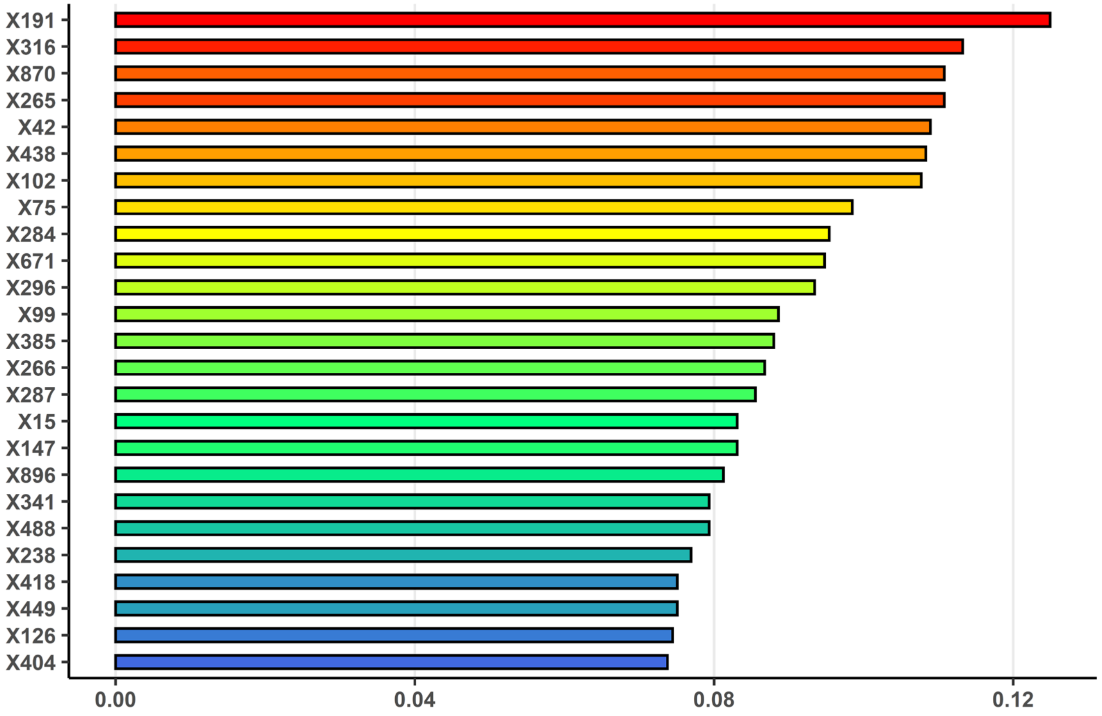
	TAP1	**TAP2**
TAP1	6	0
TAP2	1	6

92% accuracy

Feature importance plot of the 25 highest scored features from random forests classification done using data from A1α group generated via ToxPi Method 2. Features are listed on the y-axis by "Feature_ID", and the x-axis is the *gini* index value assigned to each feature (unitless).

Random Forests Feature Importance (RFI) Results

(Top 25 Features from All Data)



RF model performance using all features

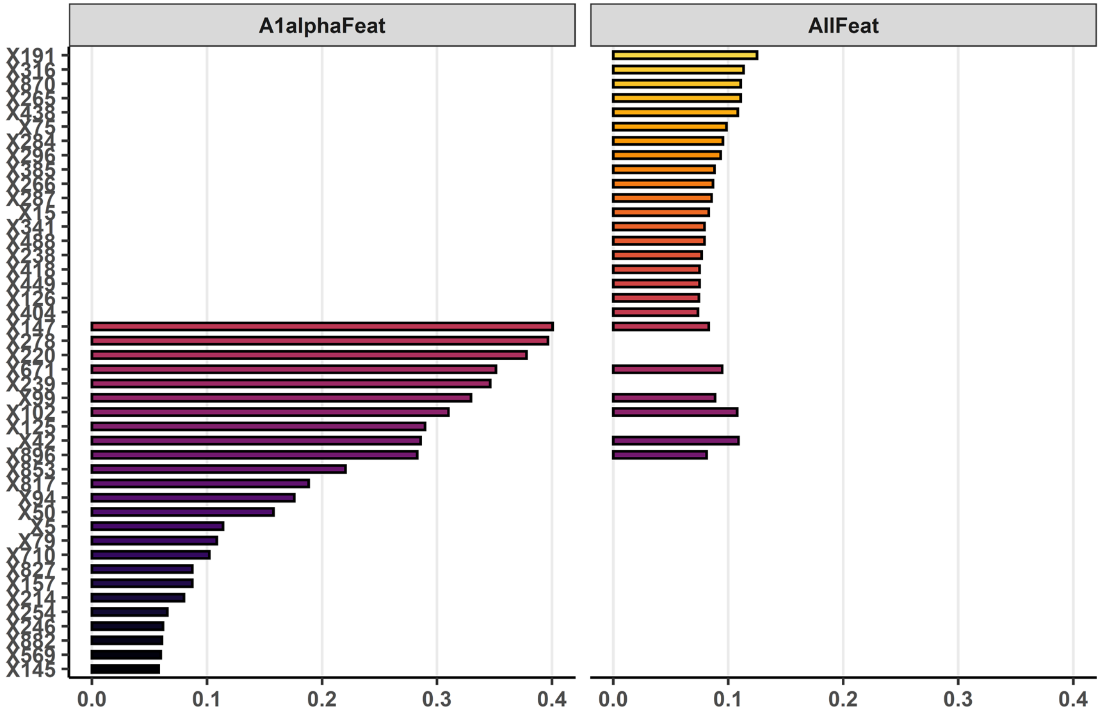
	TAP1	TAP2
TAP1	6	0
TAP2	0	7

100% accuracy

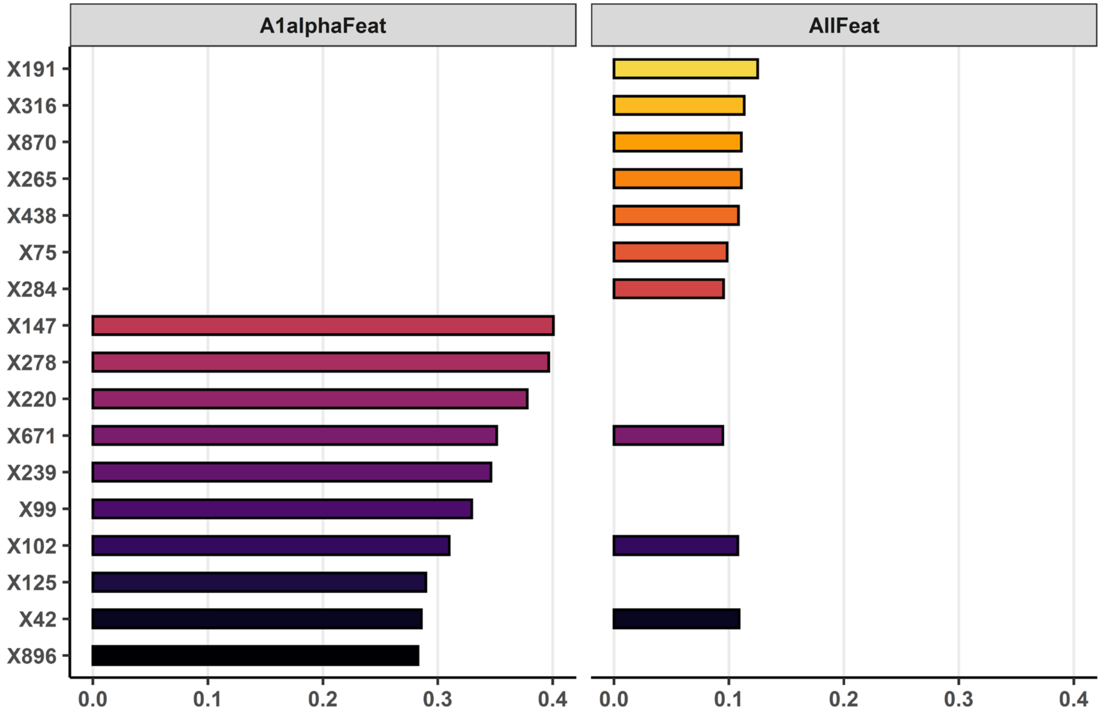
Feature importance plots of the 25 highest scored features from random forests classification done using all unique features extracted from the data. Features are listed on the y-axis by “Feature_ID”, and the x-axis is the *gini* index value assigned to each feature (unitless).

RFI Comparison: A1α Features vs. All Features

(Top 25 and Top 10)



Only 6 of 25 features in top 25 were similar between "A1α" and "all features" RFI



Only 3 of 10 features in top 10 were similar between "A1α" and "all features" RFI

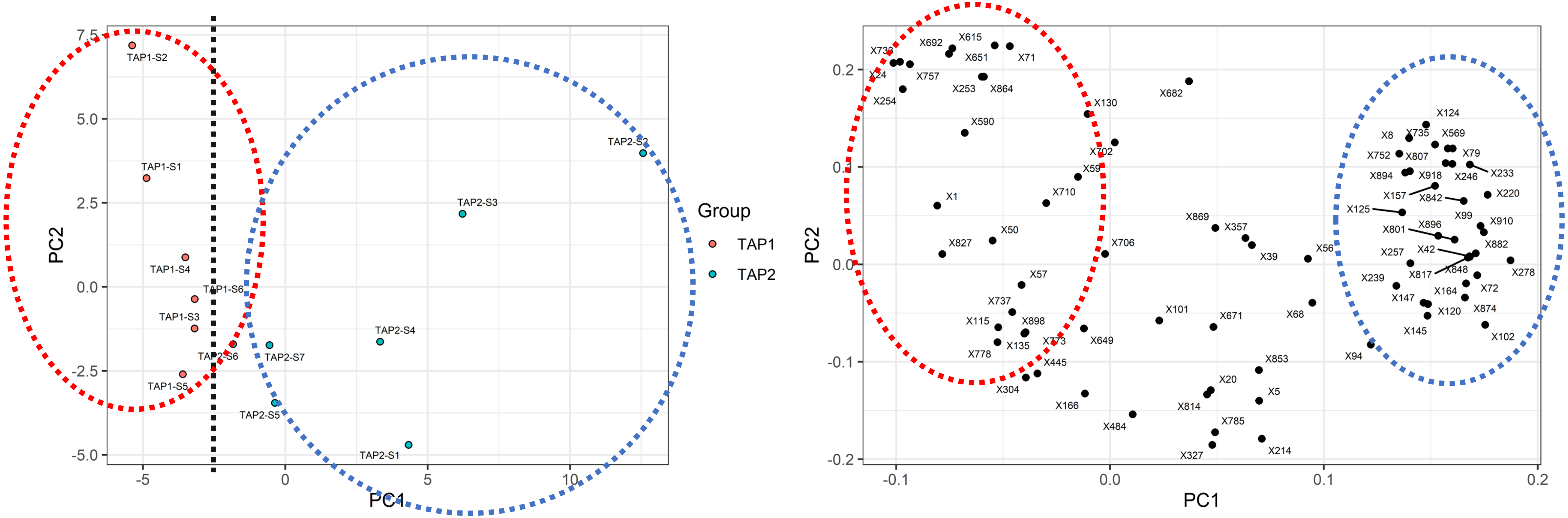
RFI: Top 10 features from All Features

(Average abundances ± standard deviations for each sample group)

Feature ID	TAP1	TAP2
X191	1583 ± 1830	171400 ± 80100
X316	236300 ± 106700	38500 ± 47470
X870	259500 ± 289800	0 ± 0
X265	0 ± 0	746100 ± 532600
X438	0 ± 0	75510 ± 37340
X75	184800 ± 67300	12840 ± 17390
X284	0 ± 0	186600 ± 157500
X671	0 ± 0	1276000 ± 953000
X102	738900 ± 614500	9331000 ± 3306000
X42	6925 ± 8256	543100 ± 360400

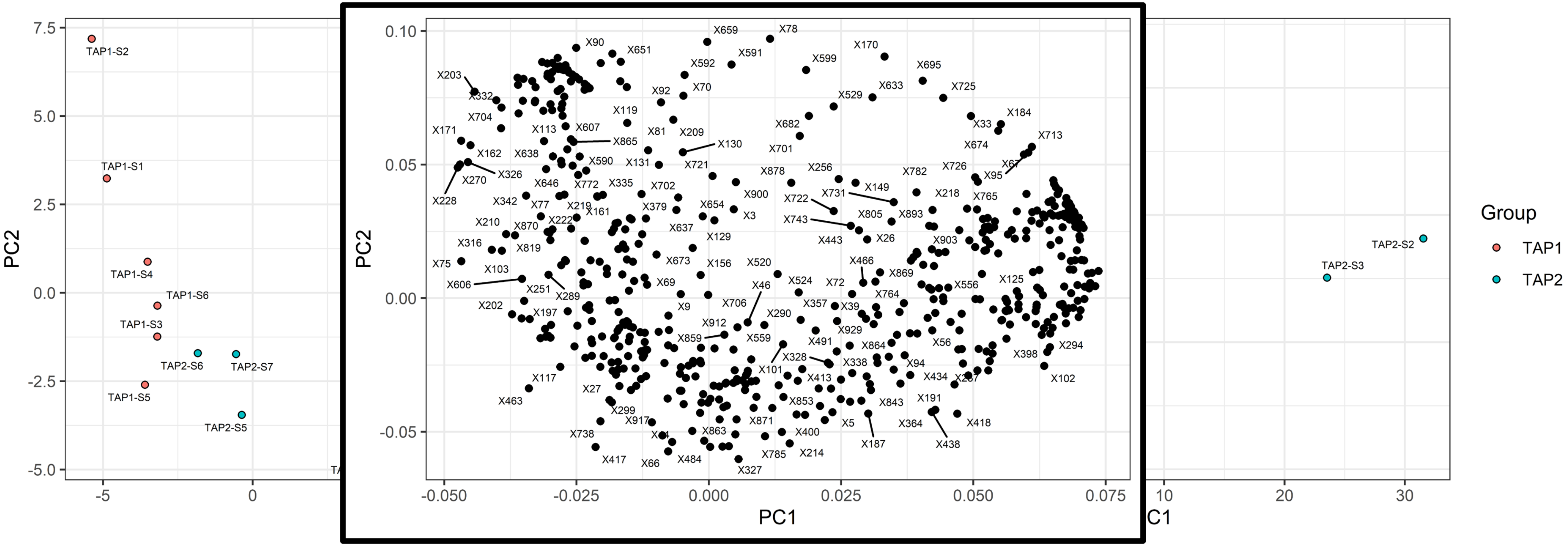
PCA Results

(A1α Features)



Scores (left) and loadings (right) plot of first two principal components (PCs) from PCA. Scores plot shows visualization of samples, and loadings plot shows features responsible for location of samples.

PCA: Comparison of A1α features vs. All features



Scores plots of PCA using only A1α features (left) and using all features (right) of first two principal components (PCs) from PCA.

Conclusions

- TEST and ToxCast method of ToxPi scoring leads to slightly different potential candidates in A1α group
    - Classification by ToxPi score and Data Source hits show which chemicals we think we have found that are “most likely” and “most interesting” based on potential harm
- PCA can separate via geographic region, RFI shows most important features at driving separation between various groups
    - Classification by RFI scoring shows which chemicals are “most important” at distinguishing one sample group from any other
- Future work:
    - Finish processing and analyzing remaining samples
    - Re-do ToxPi scores and multivariate stats workflows using the complete set of data
    - Begin determining which features are most important for further investigation

The end

- Questions or discussions?