LA-UR-22-29731

Approved for public release; distribution is unlimited.

Title:	Leveraging GCxGC-TOFMS to explore plant-soil-microbiome interactions in Maize
Author(s):	Mitchell, Joshua Merritt
Intended for:	This is a collection of slides that I can show at a job interview. No such interview is currently scheduled but since I'll leave LANL before I job search, this is the best way to take it with me.
Issued:	2022-09-20

Los Alamos NATIONAL LABORATORY

LANL Work Presentation

Leveraging GCxGC-TOFMS to explore plant-soil-microbiome interactions in Maize

Joshua Mitchell, PhD, MD

Sept 13, 2022

INSERT LA_UR HERE

Scientific Contributions and Significance to LANL

1. Technical Training and Specialties
   - What important skills and specialties do I bring to the Group and LANL as a whole?
2. Mission Related Research
   - How does my research address the mission of the Group/LANL?
3. Contributions to the Mission & Program Development
   - What are my objectives for the remaining duration of my appointment?
   - How will I contribute to program development?
4. Goals for rest of time at LANL

1. Technical Training and Specialties

B.S. Chemistry (University of Louisville, 2012)

Ph.D. in Biochemistry (University of Kentucky, 2019)

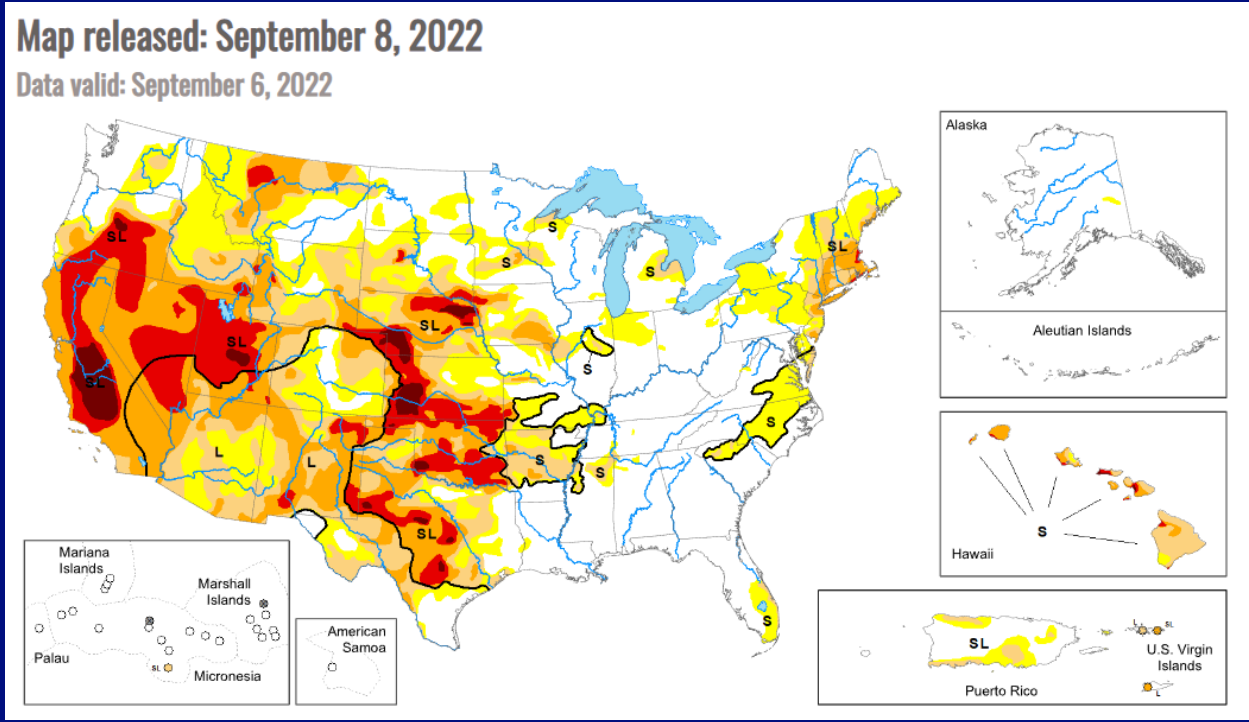
- Advisor: Dr. Hunter Moseley
- Thesis: Computational Tools for the Untargeted Assignment of FT-MS Metabolomics Datasets.
- Minor: Unofficially – Computer Science / Engineering

MD (University of Kentucky, 2021)

- Focus in Family Medicine / Primary Care and Non-Small Cell Lung Carcinoma Metabolomics

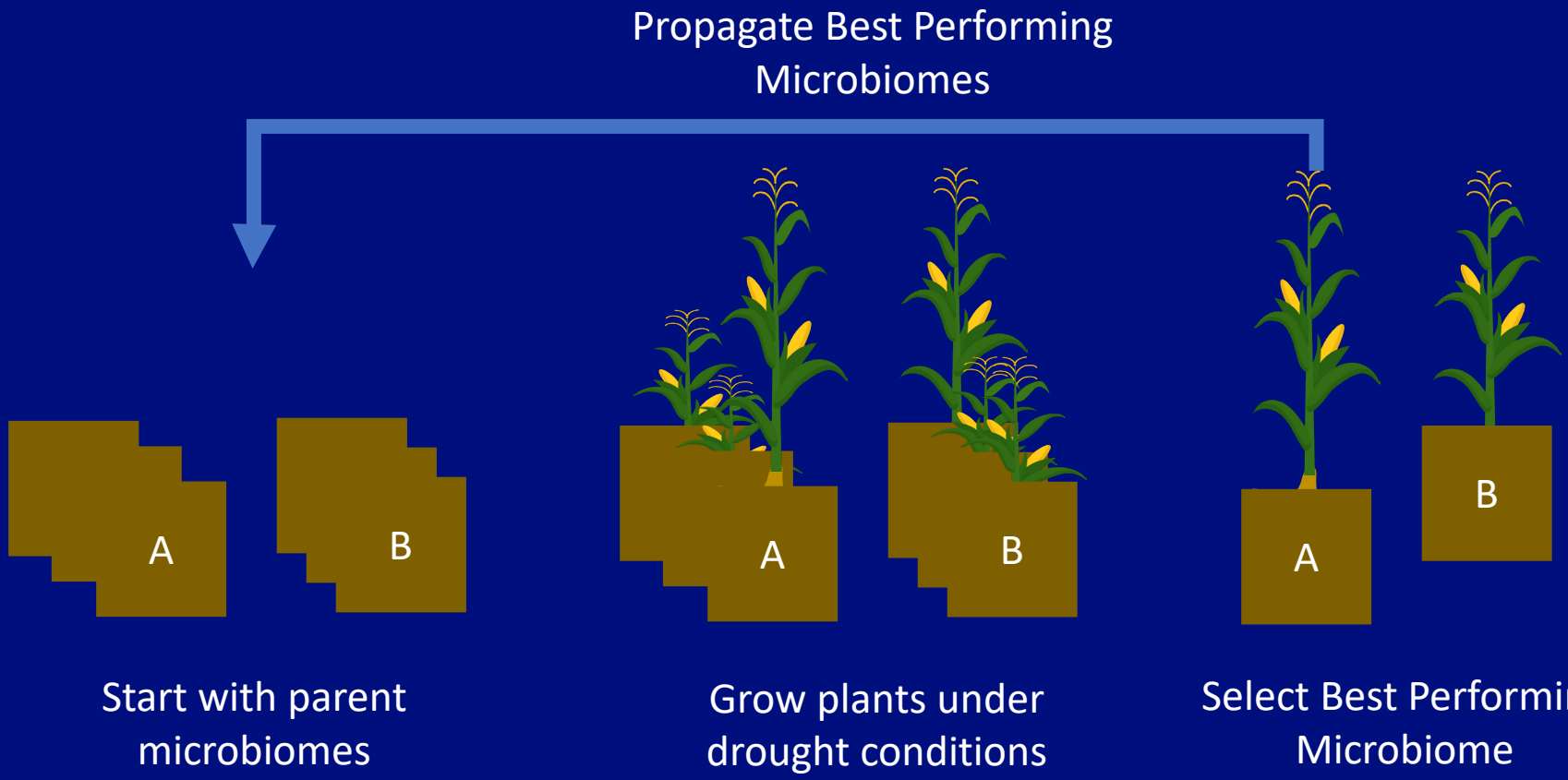
Unique Skills: My research has focused on improving our understanding of the Maize-soil-microbiome in order to direct interventions to improve crop yield. I've applied my background in computer science and metabolomics to use GCxGC-TOFMS to explore these systems and develop computational tools to identify unknown metabolites.

Motivation

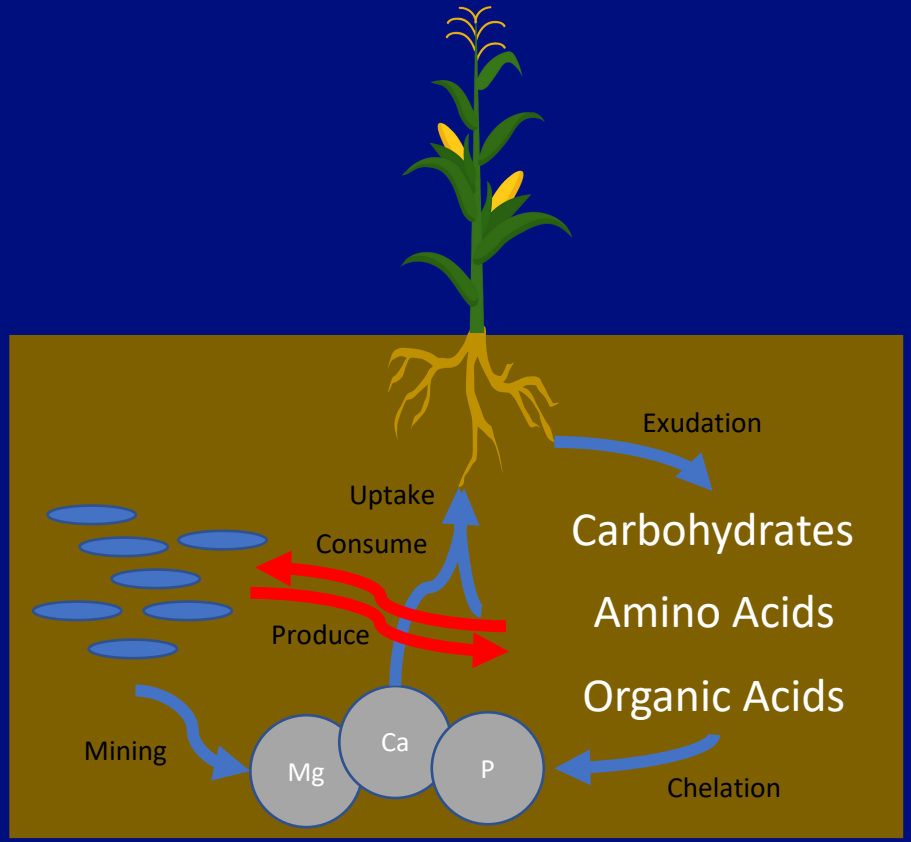


Drought (plus population growth) threatens food security in the US and the world at large.
Large-scale irrigation is impractical, is there an alternative?

Overall Experiment



But why do some microbiomes perform better?



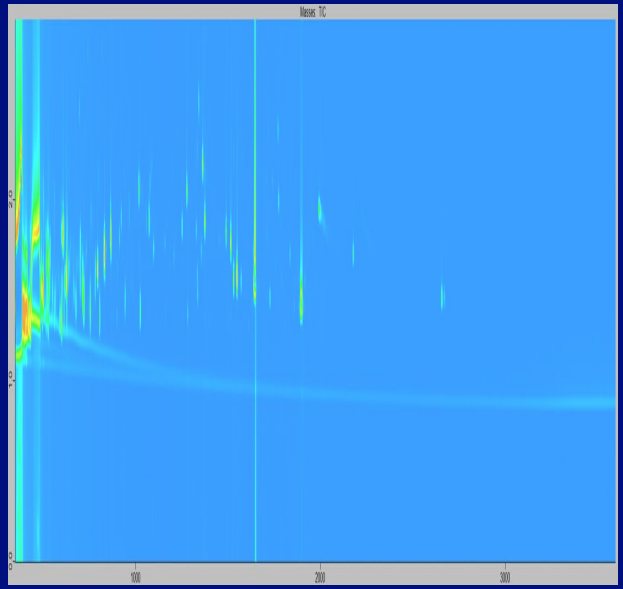
Soil Microbiomes and Plants work together to mine for nutrients and provide for one another metabolically.

Signaling and small molecules are exchanged between plants and microbes that could improve plant growth or drought tolerance.

Need to explore the root metabolome to understand further!

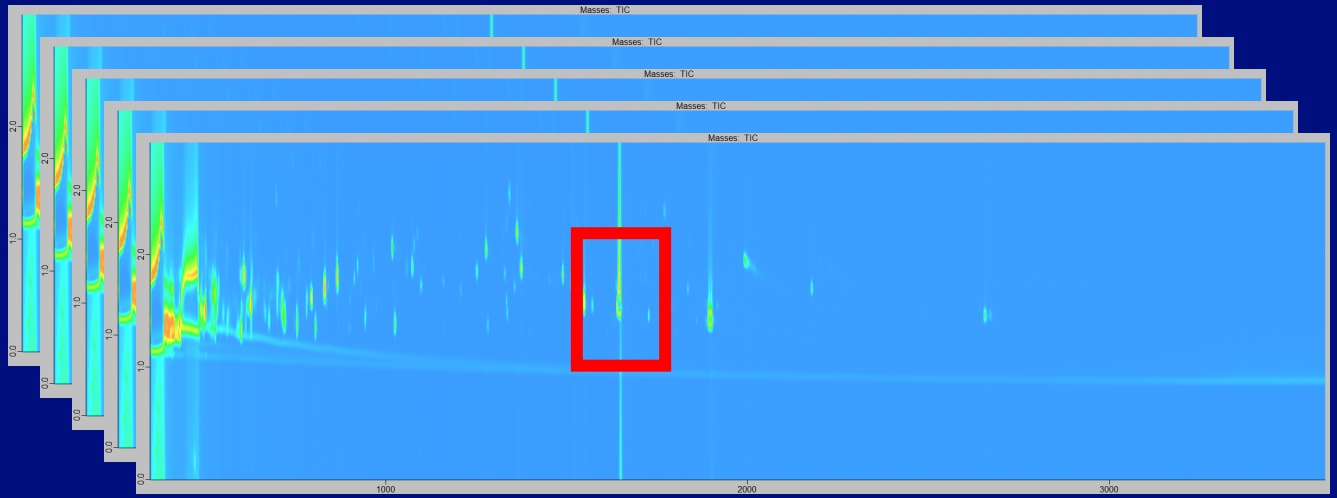
Instrumentation

LECO Pegaus GCxGC-TOFMS System



An example 2D chromatogram. Each pixel is a mass spectrum.

Analysis Tools



Tile-Based Fischer Ratio Analysis (ChromaTof Tile) was employed to identify differentially abundant chemical features between classes.

Rules based on S/N ratios, minimum number of significant mass channels, and feature occurrence in at least N samples were used to minimize false positives.

The parent microbiomes

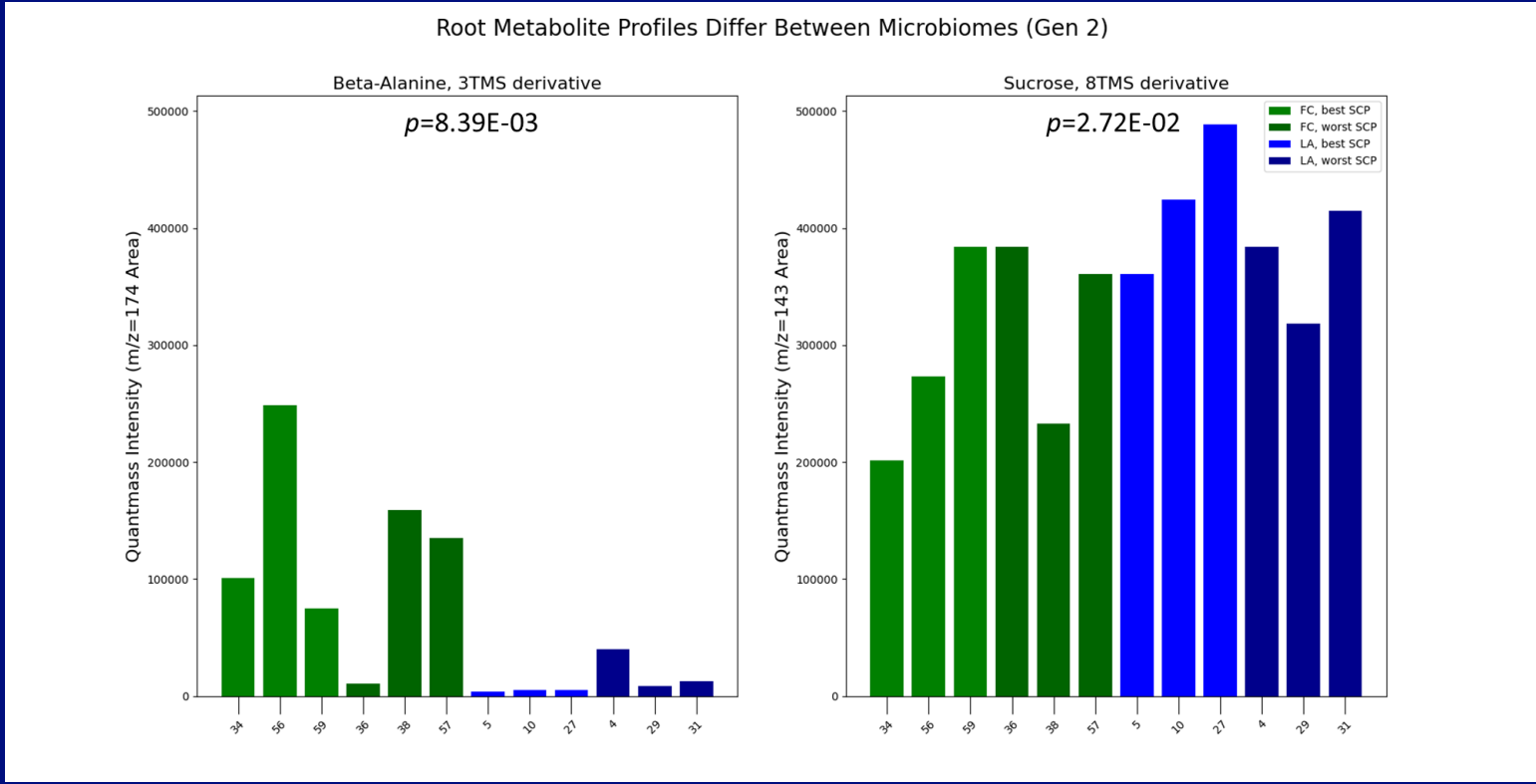
A

A: Los Alamos Forest Microbiome

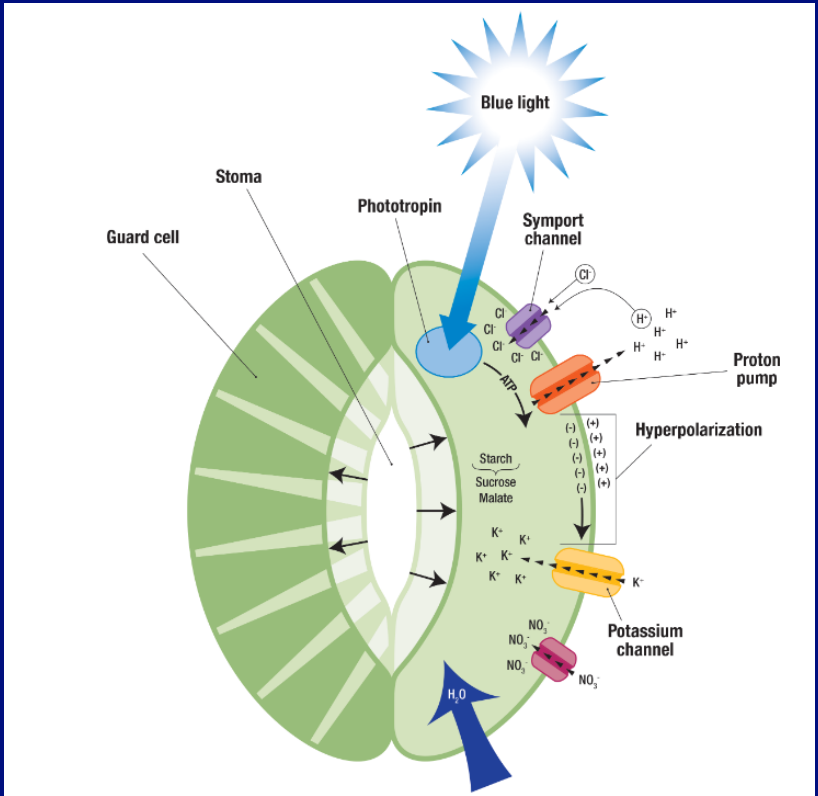
B

B: Fort Collins Agricultural Field Soil Microbiome

Parent microbiomes induce different metabolite responses in maize roots.



What is stomatal closure point?



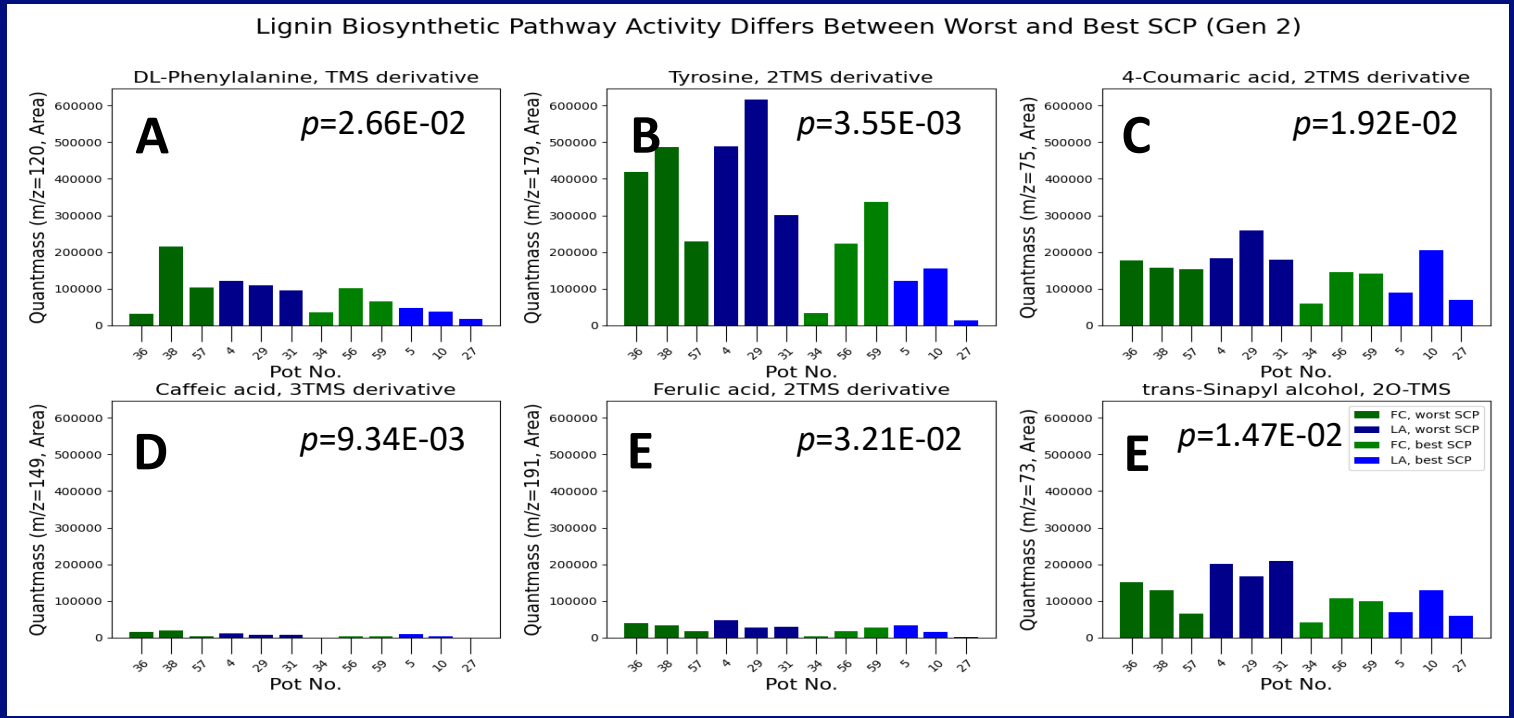
Open stomata enable gas change but result in water loss.

Photosynthesis needs open stomata.

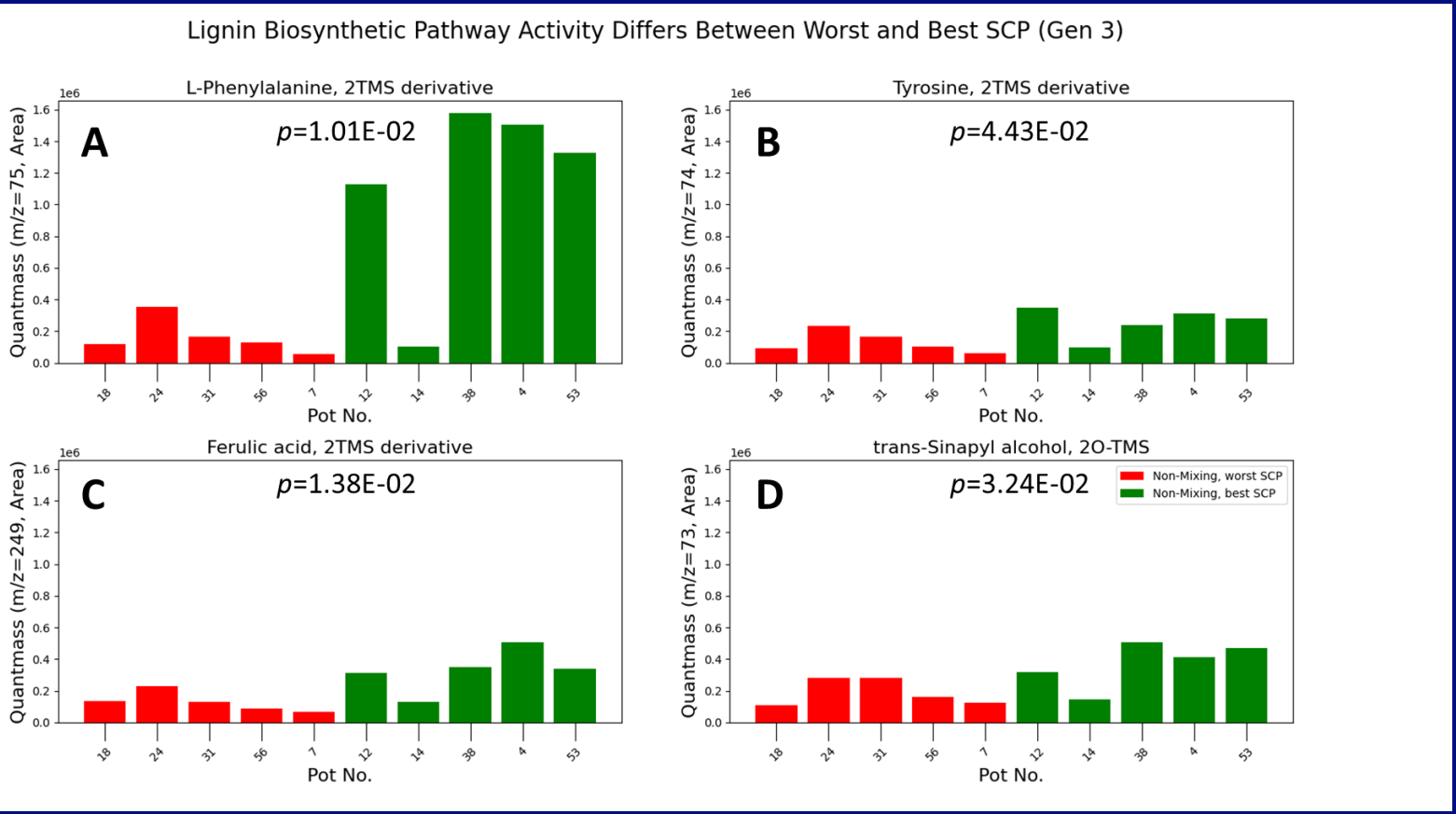
The decision to open stomata requires balancing energy production with water loss.

A ‘worse’ SCP means more often closed stomata, a ‘better’ SCP means more often open stomata.

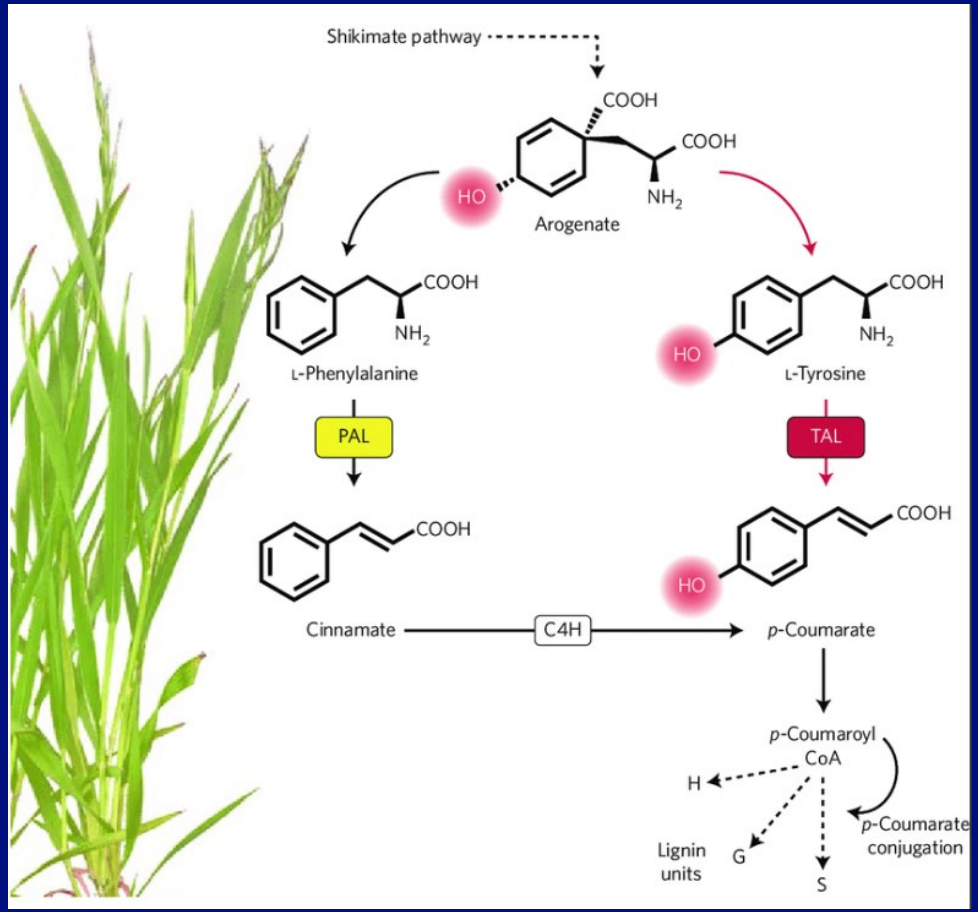
Generation 2 roots show differences in lignin biosynthesis depending on stomatal closure point



But generation 3 roots have the opposite pattern!



What is Lignin?



So what does this all mean?

The agricultural field soil results in higher beta-alanine levels in the roots. The plants are reacting as if the drought is more severe than it actually is.

The forest soil results in plants producing more sucrose, perhaps to recruit beneficial microbes.

In the generation where a statistically significant difference in SCP was observed, the lignin pathway appears **more active** in the worst SCP plants, the opposite of generation 3.

Since the plants from generation 2 to generation 3 were from the same seed source (i.e., basically clones) this is probably due to the microbiome.

Lignin is a structural carbohydrate associated with growth and drought tolerance but many questions remain.

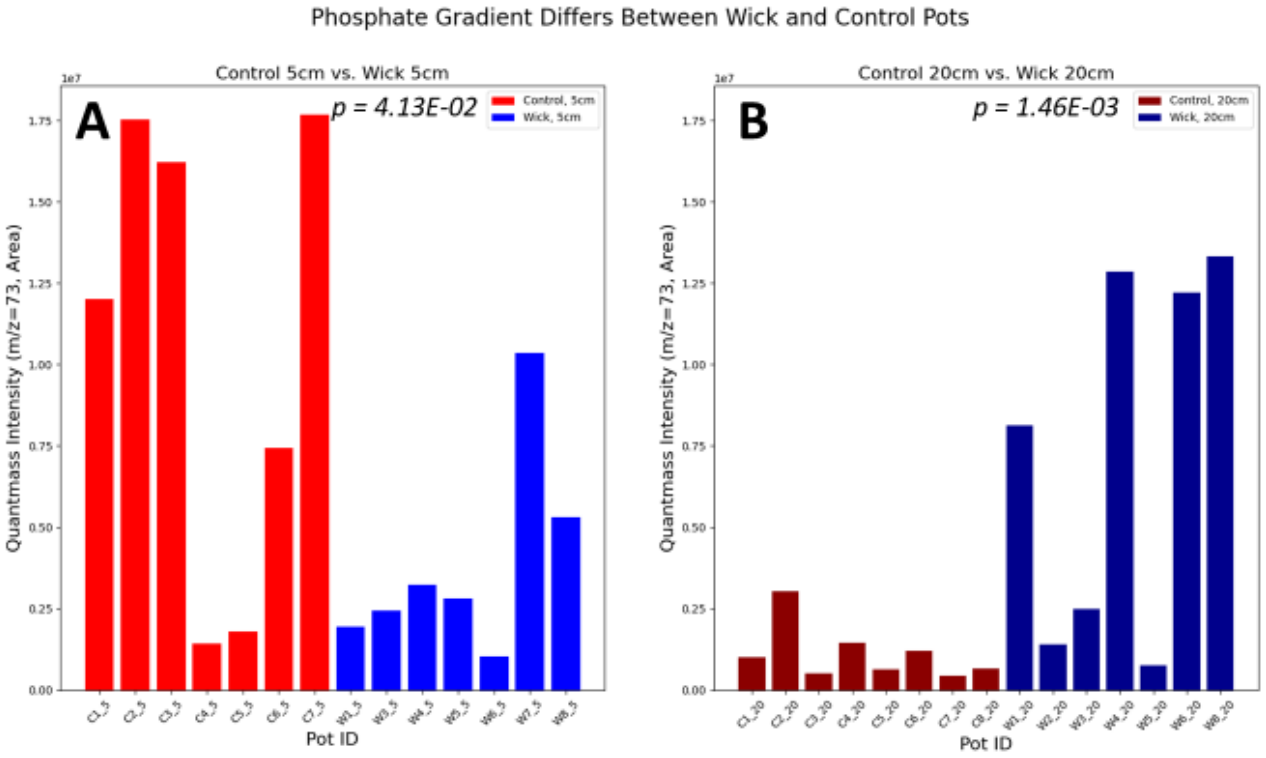
Side Experiments – Wick vs. Standard Pots

!=

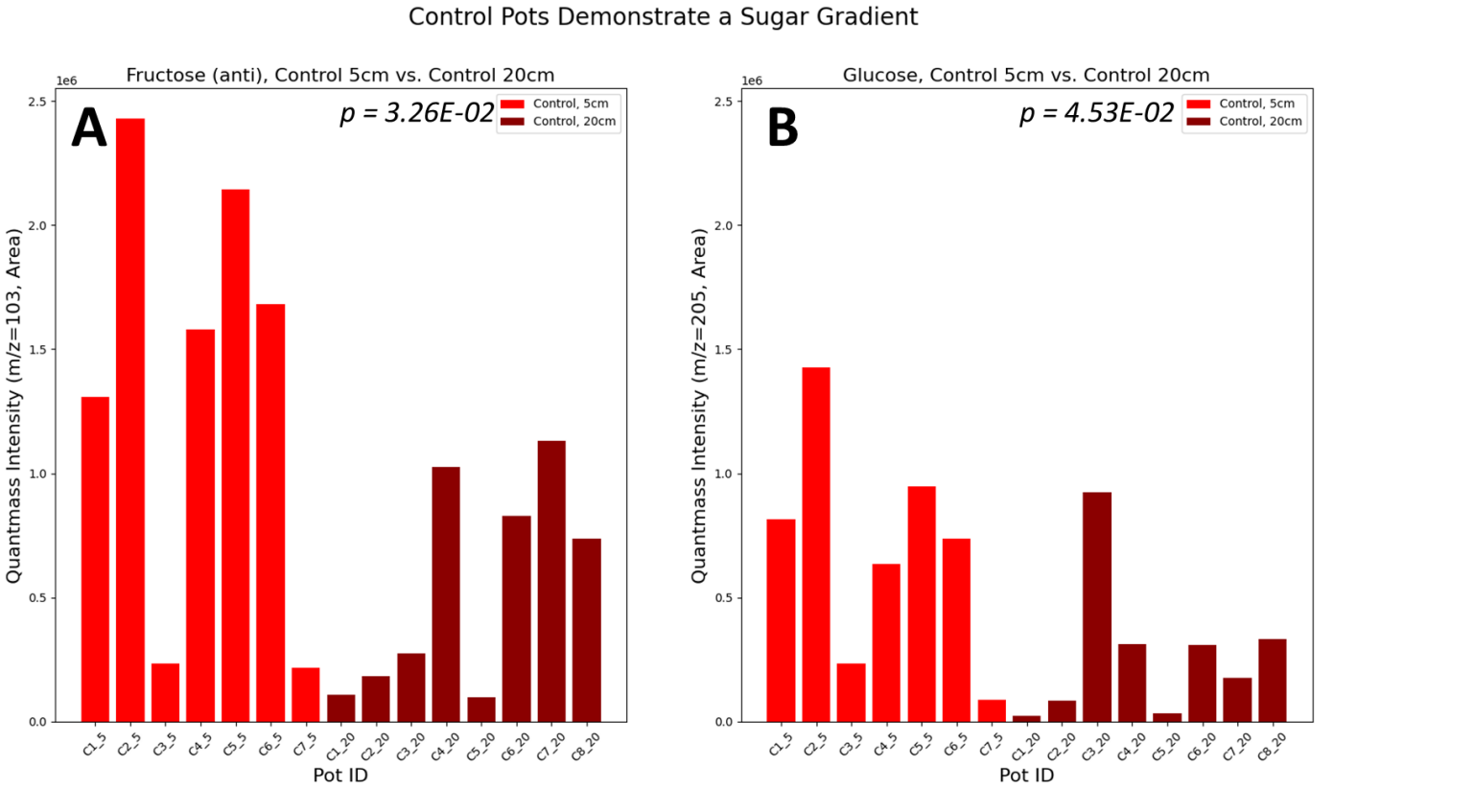
Pots are not fields, but many experiments are performed using pots for convenience. Notably pots have a bottom, fields do not, changing the hydrology.

Can the addition of a wick to a pot change the distribution of compounds by mimicking the suction force seen in real fields?

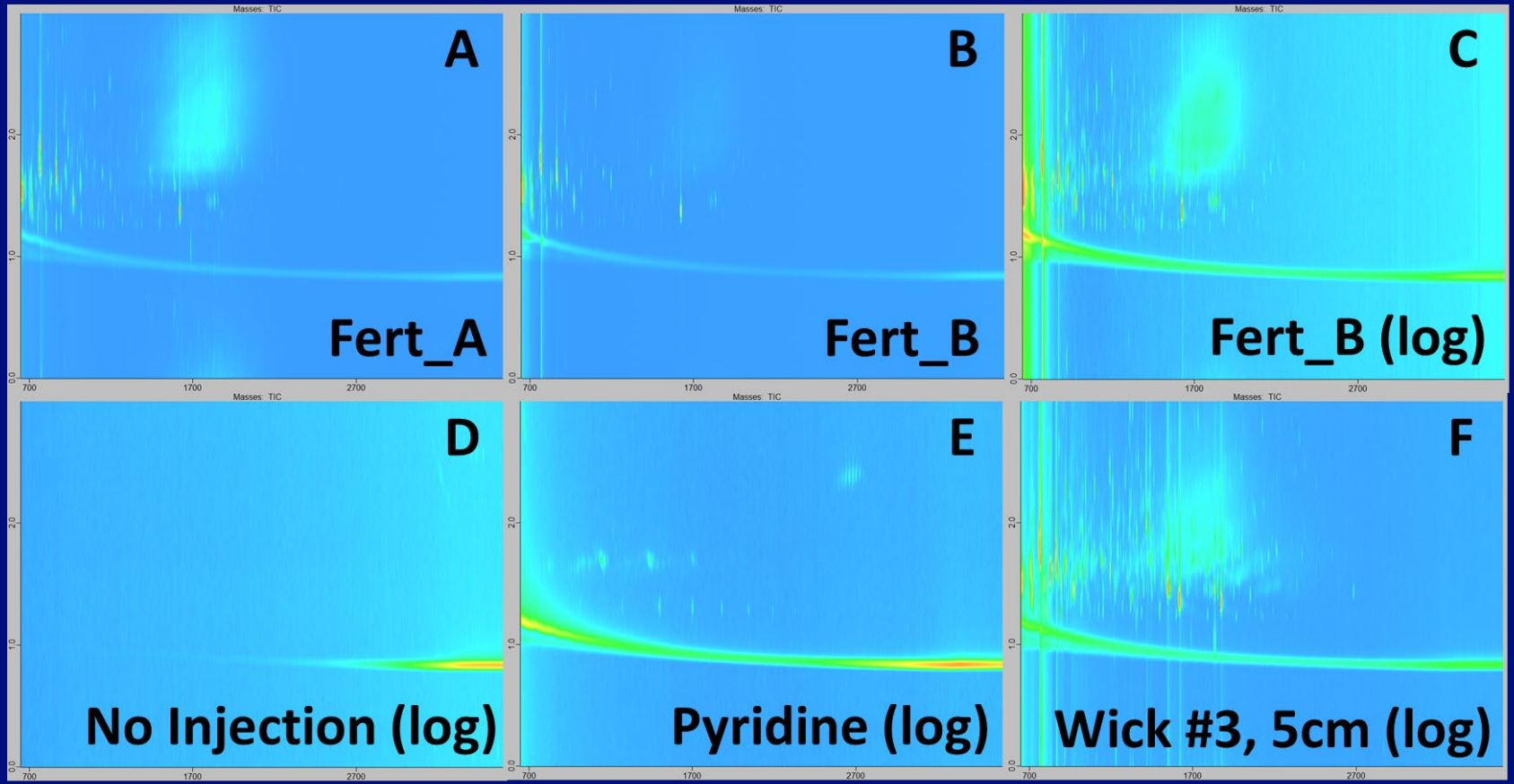
We see differences in fertilizer distribution



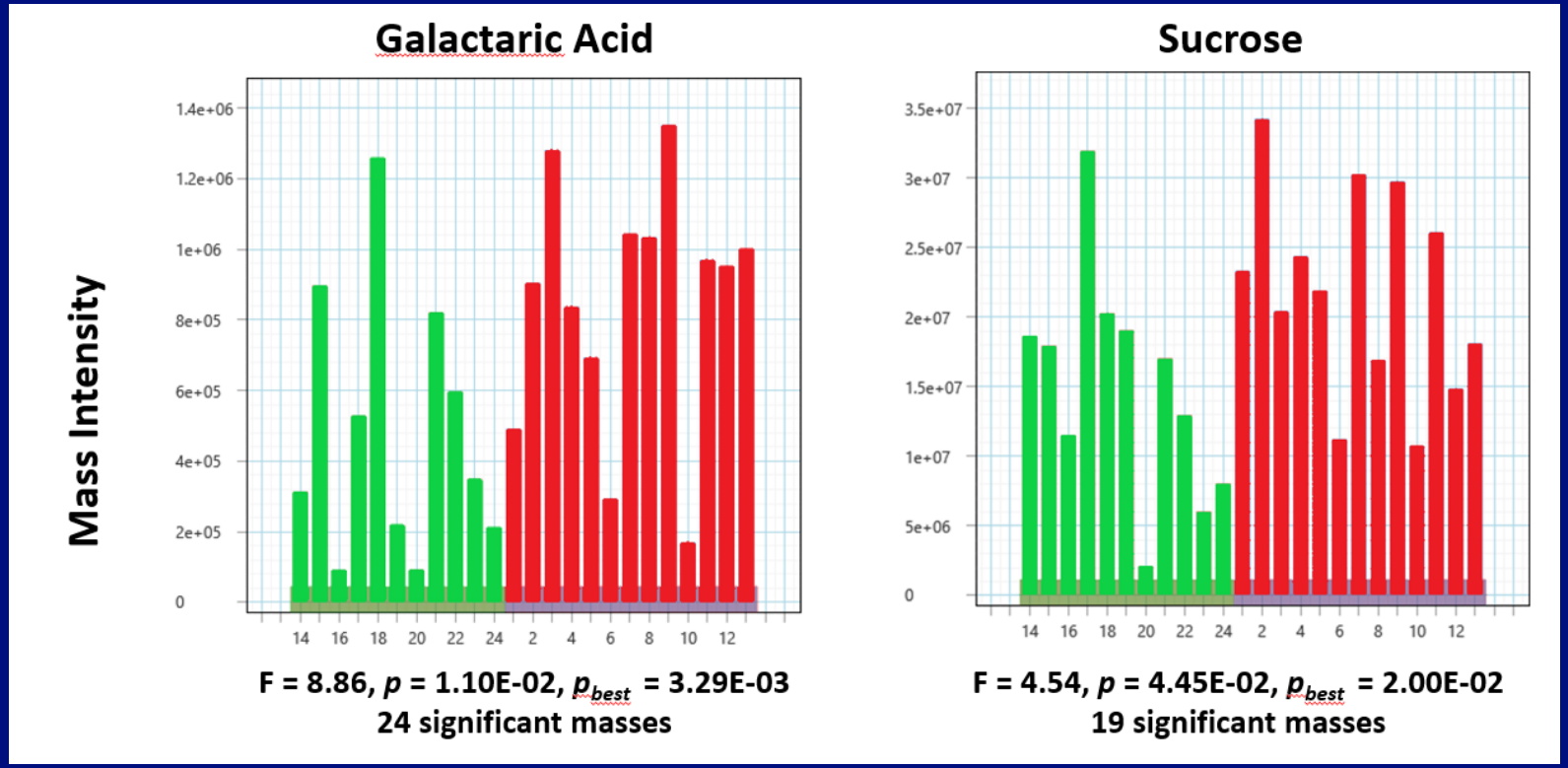
We see differences and in sugar gradients



But salty aqueous solutions are less than ideal for GC-MS

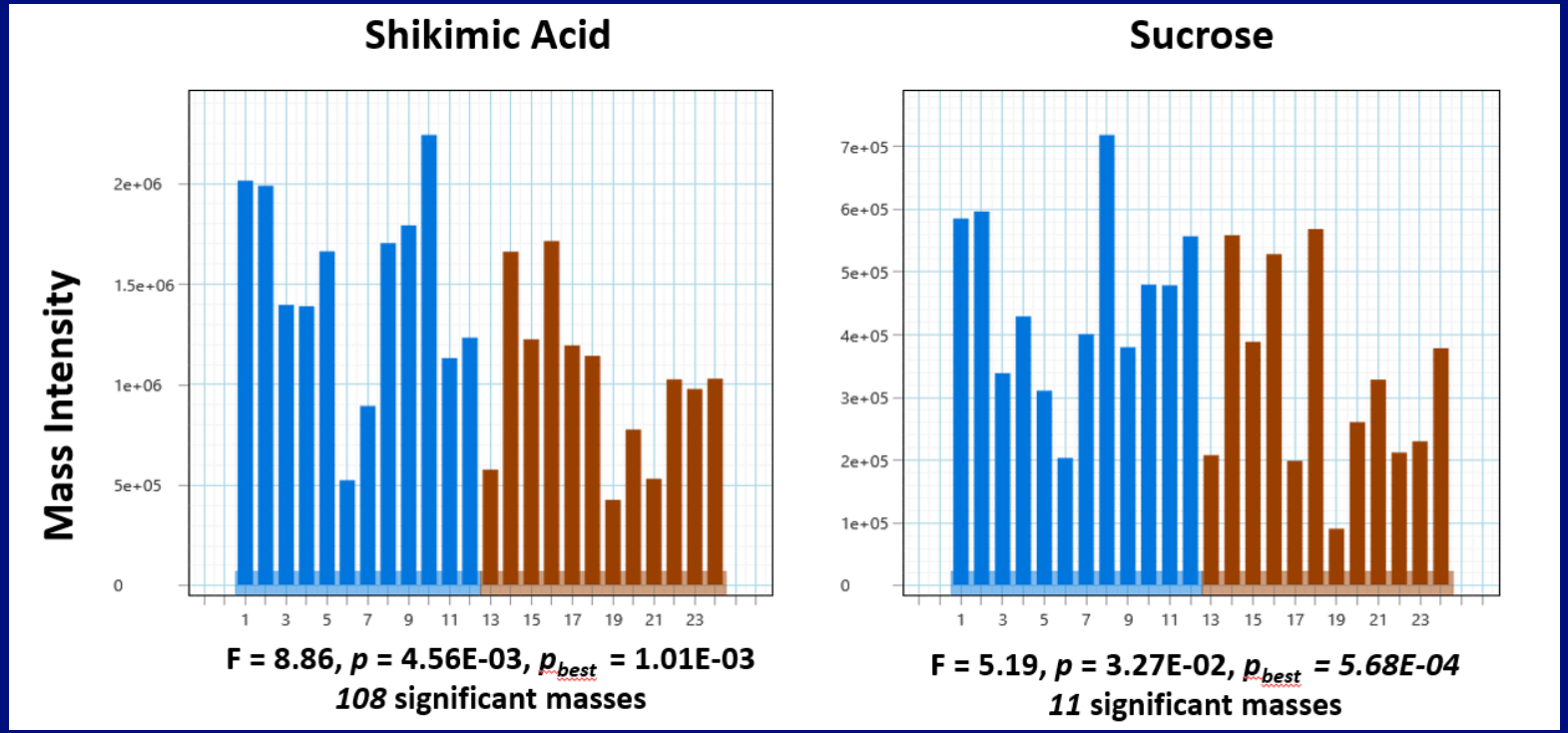


Side Experiments – Uninoculated Pots x Drought



The absence of a microbiome in the soil causes Maize roots to produce more sugar and organic acids.

Side Experiments – Uninoculated Pots x Drought



Drought induces changes in core metabolic pathways.
They appear less abundant in drought.

Side Experiments – Exudate Composition

Our Quantification

Compound	un-concentrated (ppm)	un-concentrated (mM)
Sucrose	0.137	0.0004003
Inositol	0.038	0.000211135
Mannitol	0.061	0.000337373
Fructose (iso 1)	0.163	0.00090643
Fructose (iso 2)	0.163	0.00090228
Glucose (iso 1)	0.359	0.001990141
Glucose (iso 2)	0.198	0.001096686
Quinic Acid	0.241	0.001256641
Shikimic Acid	0.020	0.000117393
Aconitic Acid	0.276	0.0015859
Ribitol	0.000	0
Ribose	0.014	9.45189E-05
Arabinose	0.009	6.13035E-05
Malic Acid	0.023	0.000169612
Succinic Acid	0.000	0
Pyruvic Acid	0.057	0.000646356
Benzoic Acid	0.000	0
Urea	0.000	0
Stearic Acid	0.013	4.59294E-05
Palmitic Acid	0.050	0.000195157
Myristic Acid	0.000	0
Glycine	0.000	0
Alanine	0.000	0
Aspartate-3TMS	0.000	0
Proline	0.000	0
Glutamine	0.000	0
Serine	0.000	0

Zhu et al.

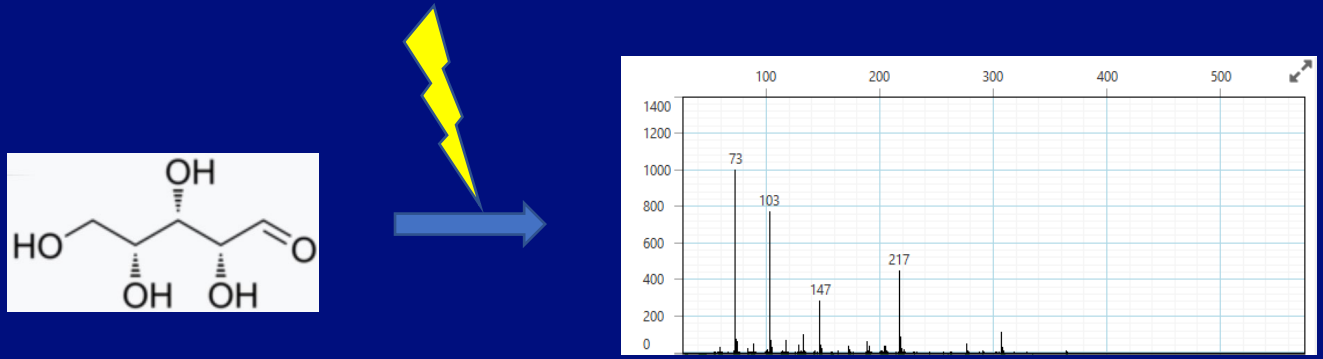
Compound	N rate			
	N0	N30	N80	N160
–	5566 b	4652 b	6433 ab	10799 a
All sugars	317 b	167 b	344 b	3459 a
6-deoxyglucose	21 b	22 b	48 b	405 a
fructose	71 ab	8 b	17 b	862 a
fucose	11 b	28 ab	56 a	53 a
galactinol	2 b	5 b	12 ab	65 a
galactose	108 b	13 c	24 c	1150 a
glucose	17 b	6 b	13 b	206 a
glucose-1-phosphate	7 b	7 b	14 ab	36 a
lactose	2 b	8 a	14 a	14 a
levoglucosan	5 b	7 b	14 ab	34 a
mannose	5 b	5 b	14 b	61 a
sucrose	14 ab	6 b	21 ab	202 a
threose	1 b	5 a	10 a	12 a
xylose	44 b	36 b	65 b	331 a
All sugar alcohols	31 b	31 b	64 b	921 a
erythritol	5 b	6 b	13 ab	35 a
lyxitol	6 b	7 b	9 b	744 a
myo-inositol	5 b	5 ab	10 ab	34 a
xylitol	12 b	6 b	21 ab	90 a
All amines	125 a	63 a	92 a	161 a
cyclohexamine	2 c	6 bc	15 b	70 a
glycine	2 b	7 ab	15 a	14 a
All carboxylic acids	382 a	742 a	1116 a	799 a
2-hydroxyvaleric acid	8 b	26 ab	44 a	61 a
4-hydroxybutyric acid	3 b	6 ab	15 a	18 a
azelaic acid	2 b	5 ab	12 a	12 a
dihydroxymalonic acid NIST	2 b	5 ab	11 a	14 a
phthalic acid	33 b	16 b	39 b	170 a
quinic acid	3 b	7 ab	12 a	13 a
All polyols	1945 a	458 b	928 ab	1595 a
glycerol-3-galactoside	2 c	6 bc	8 b	43 a
phenylethanol NIST	4 b	11 ab	34 a	21 a
stigmasterol	1 c	6 b	17 a	19 a
All phenolics	348 b	1031 ab	2165 a	2327 a
hydroquinone	3 b	17 ab	31 a	15 ab
p-cresol	306 b	924 ab	2013 a	2199 a
All lipids	519 a	1115 a	1076 a	1312 a
1-monopalmitin	13 b	25 ab	29 ab	35 a
2-deoxyerythritol	8 b	10 b	17 ab	38 a
2-monopalmitin	6 b	24 b	59 ab	73 a
caprylic acid	6 b	15 ab	25 a	20 a
palmitoleic acid	2 b	6 a	12 a	12 a

Significant differences to established literature.
Highlights the problem with commonly used metabolite assignment methods

Main Experiment – Future Directions

- Instrument Issues:
  - The LECO is finicky with tuning and intensity drift.
  - Could explore this with controlled experiments
- Main experiment results:
  - Lignin pathway activation could be further examined:
    - Transcriptomics
    - Stable-isotope tracing
    - Microscopy
  - More investigation of what happened between gen 2 and gen 3.
    - LC-MS + FT-MS to capture more metabolites
    - Transcriptomics
    - Soil microbiome metagenomics
  - More replicates
    - Basically this is an n=1. An interesting n=1 but still n=1.

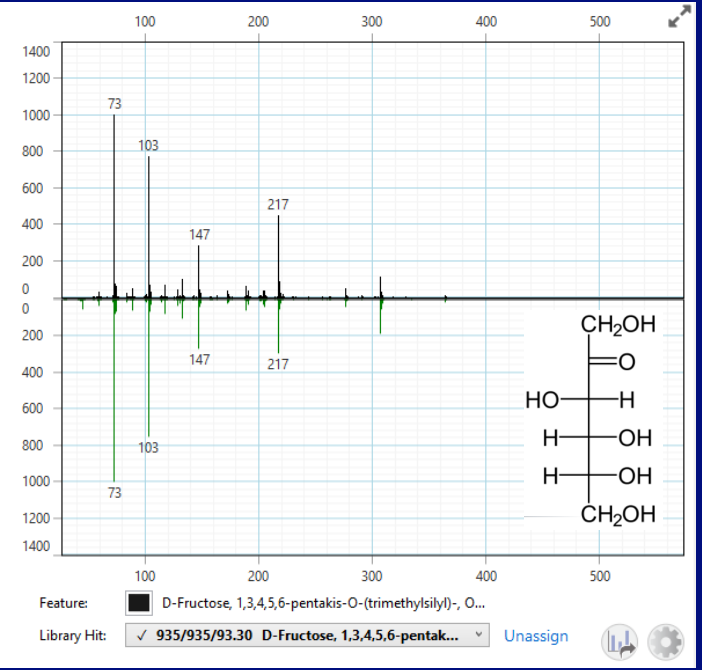
Side Project – Machine Learning Pipeline to Improve Assignment Accuracy

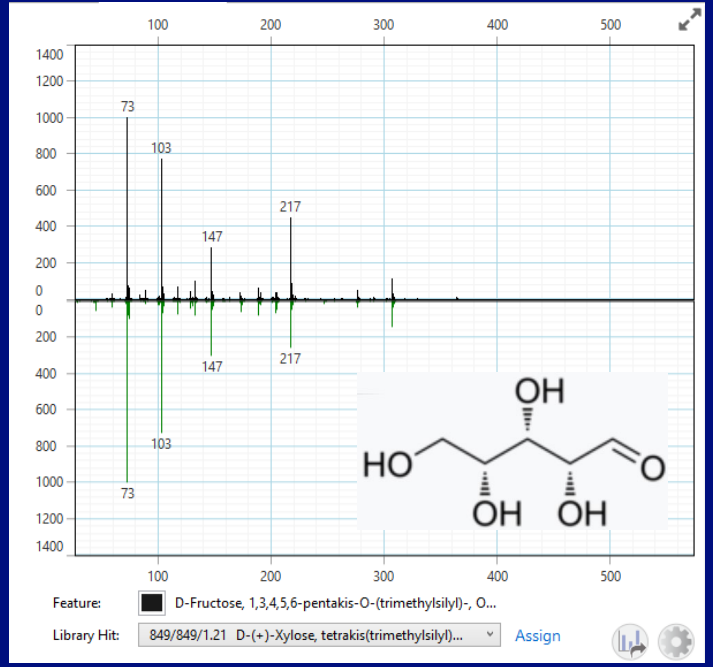


By blasting a chemical structure with an electron beam and fragmenting it, a mass spectrum can be produced.

Each peak is the mass-to-charge ratio of a fragment ion.

Side Project – Similar Structures = Similar Spectra





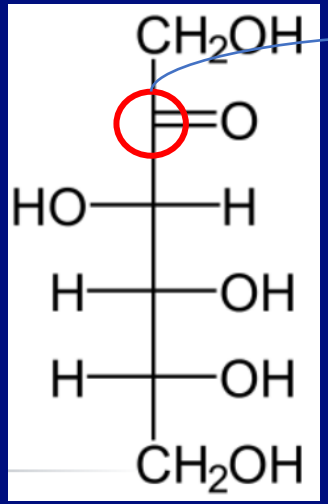
Is this Fructose or Xylose.

Very different sugars biologically!

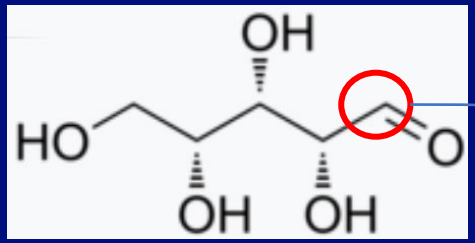
Structural similarity = Mass spectral similarity

But one is *keto* sugar, the other *aldo*…

Side Project – Similar Structures = Similar Spectra



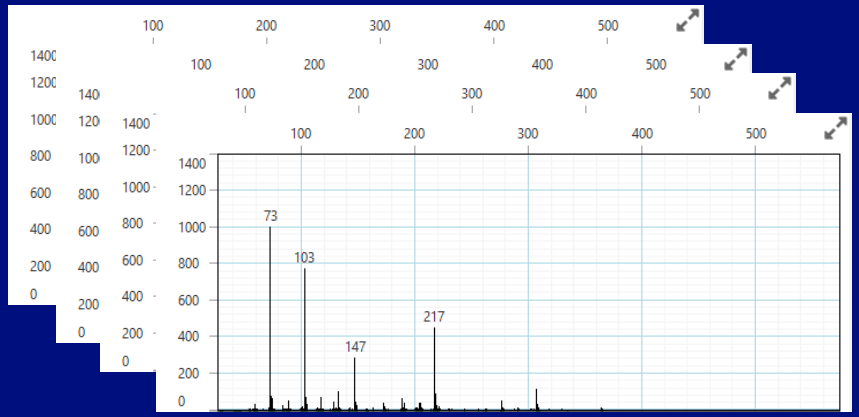
D=0, C
D=1, C((C,1),(C,1),(O,2))
D=2,
C((C((C,1),(O,1),1),(C((C,1),(C,1),(O,1),1),(O(C,2),2))



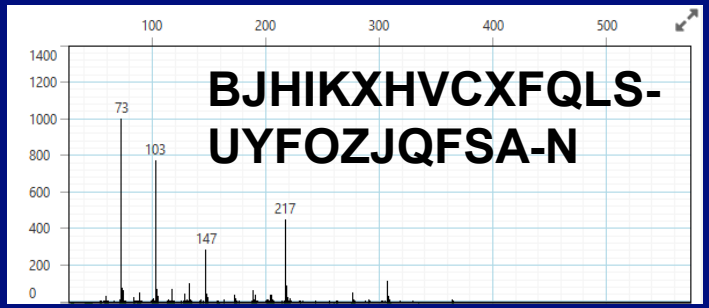
D=0, C
D=1, C((C,1), (O,2))
D=2,
C(C((C,1),(C,1),(O,1),1),(O(C,2),2)

If we can predict the presence of a ketone or aldehyde we can discriminate between Fructose and Xylose.
“Graph coloring” enables the efficient representation of any substructure.
Can we predict these colors using machine learning?

Side Project – Training Data



~300k EI-MS Spectra from NIST



But entries do not have structures!
Just InChiKeys…

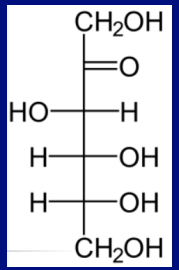
BJHIKXHVCXFQLS-UYFOZJQFSA-N

PubChem API

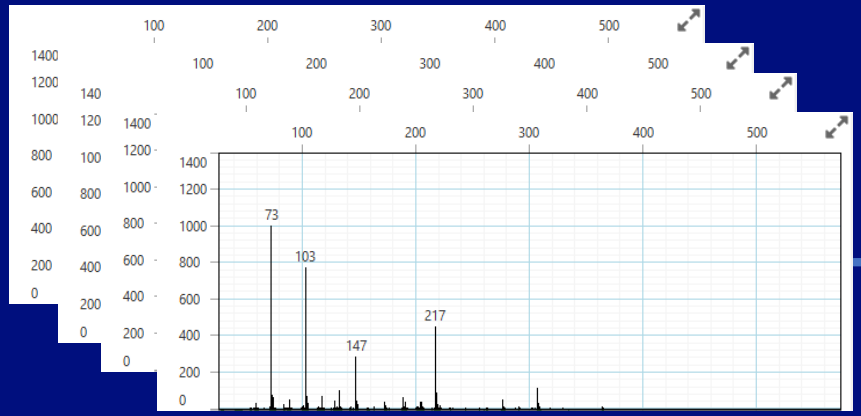
Local SQLite Caching

C1C(C(C(C(O1)(CO)O)O)O)O

PySMILES



Side Project – Deduplicate



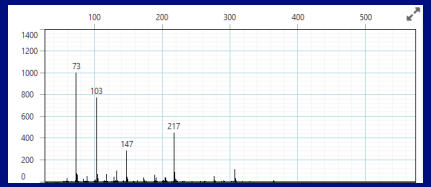
Deduplicate Entries Based on SMILES Strings

Hold Out Set

Training Set

Side Project – Automated Feature Engineering & Binary Color Models

Training Set



Feature Vectors

[0,0,0,123,4,0,0…]
A length 500 vector of ints

Engineered Vectors

Use FWE selection to drop uninformative features

Binary Model per Color

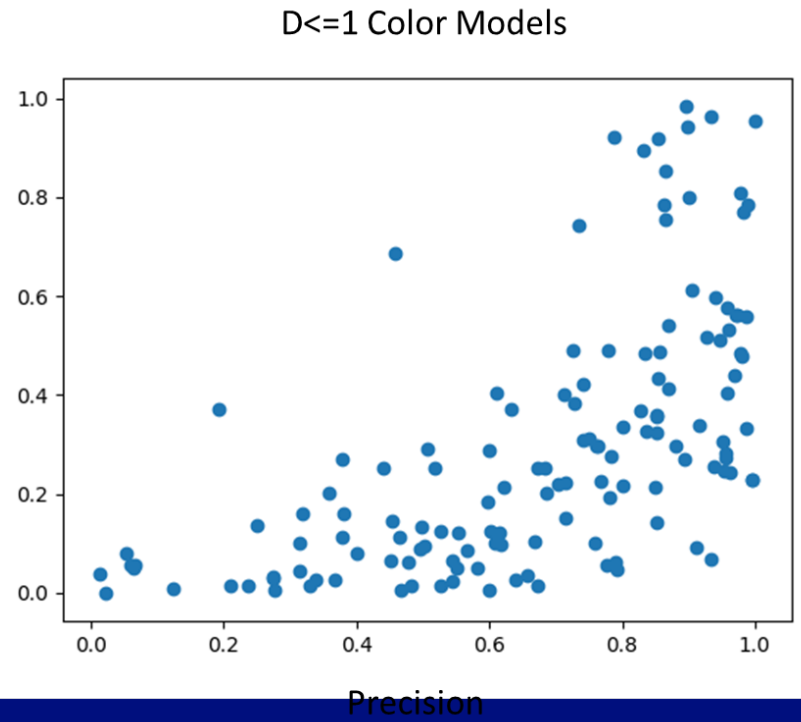
Use RandomForest to build a model using engineered vector

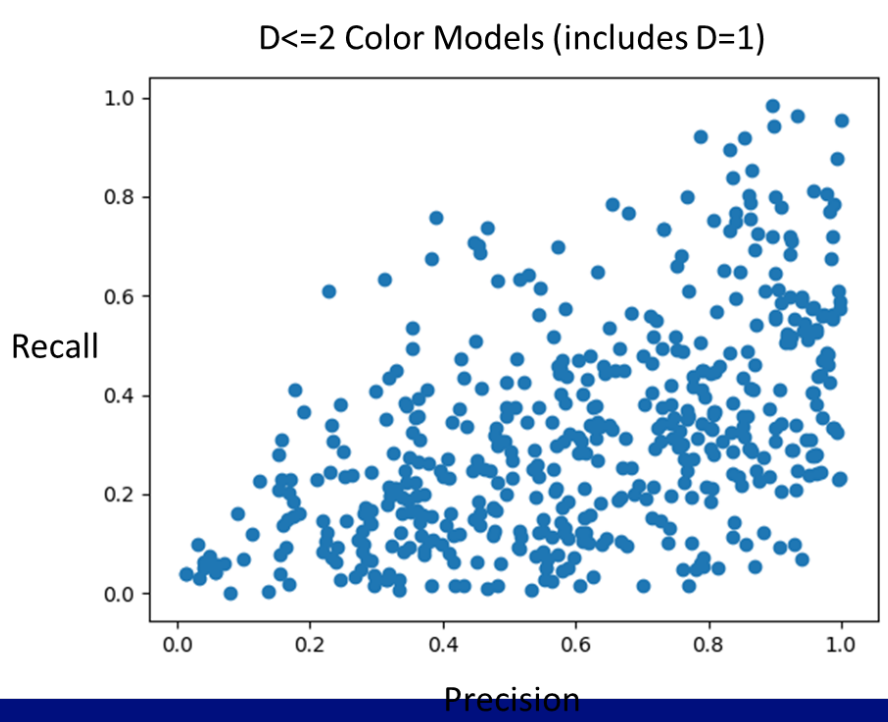
Side Project – Model Storage

Store models and FWE selector using JSONPickle

…
{
“Color”: “C((C,1),(C,1),(O,2))”,
“Count”: 120345,
“Transformer”: pickled FWE selector,
“Model”: pickled RF model,
“Hyperparam Dict”: …,
“Performance”: { *from 5-fold stratified cross validation*
“MCC”: .89,
“Precision”: .99,
“Accuracy”: .99,
“Recall”: .37,
}
}

Side Project – Models tend to be high precision and low recall





Side Project – Combine Models

Ubiquitous colors are uninformative
Rare colors are uninformative

Which models need to be included in a model?

- Doing this by hand is bound to be error-prone

How do we evaluate a model combination?

- Can we calculate a weighting based on known substructures in a possible assignment and predicted substructures?
- Combined with cosine similarity, a traditional metric for comparing mass spectra.
- If the weighting improves the net or average **rank** of the correct assignment, then it's a good combination.
- Can evaluate using the hold-out training dataset.

Side Project – Combining with PyGad

A model combination can be expressed using a bitmask.

- Assign each model to an arbitrary index *i.*
- If bitmask[*i*] = 1, include model; else exclude model.

The bitmask can be optimized using genetic algorithms.

- For d=1, |bitmask| = 138, i.e., 3.48E41 combinations
- For d=2, |bitmask| = 502, i.e., 1.31E151 combinations!

Needs incredible performance to make tractable.

- Pre-compute all cosine similarities (big memory footprint).
- Vectorize cosine weighting and ranking functions.
- Use Pandas and Numpy where possible.
- Use high-performance computers
    - SLURM and tricks to side step time limits on the cluster

Side Project – Wrapper Trick


```
],
"last_best_solution_fitness": -266.0,
"mutation_percent_genes": "default",
"mutation_probability": 0.1,
"mutation_type": "random",
"num_generations": 2,
"num_genes": 500,
"num_parents_mating": 5,
"parent_selection_type": "sss",
"previous_ga_instance": null,
"previous_population": [
    [
```

```
],
"query_df_path": "/vast/home/jmmitchell/query_dump_c2.pkl",
"reference_df_path": "/vast/home/jmmitchell/ref_dump_c2.pkl",
"total_generations": 96,
"total_pop": 3840,
"wrapper_path": "/vast/home/jmmitchell/wrapper_dump_9_12_22_c2_d30.json"
```


Each iteration takes the wrapper which stores all needed metadata and previous solutions.

Each generation this is updated, and a new job initiated taking the wrapper.

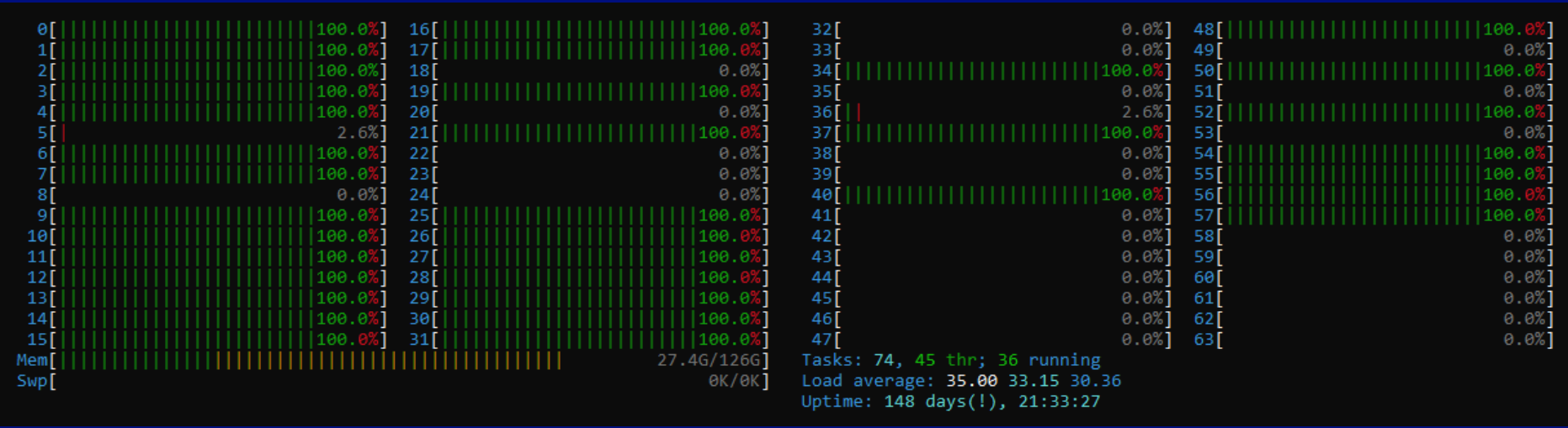
This means the job can start, stop on a whim. No need to micromanage starting and stopping to get around time limits on cluster.

Process stops when convergence or generation limits met.

Side Project – Example Solutions


```
{
    "all_solutions": {
        "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1
, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]": [
            "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0,
 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0,
0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]",
            226.0,
            2
        ],
        "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1
, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]": [
            "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0,
 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0,
0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]",
            226.0,
            1
        ],
        "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1
, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]": [
            "[0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0,
0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1]",
            228.0,
            1
        ],
```


Side Project – Enough Performance?



Evaluating 100 generation of 40 population members ~2 days of compute

100 generations sufficient for D=1 convergence but not D=2

Side Project – Results

- D=1 combinations yield a notable improvement ~300 net improved assignments out of 8000 test cases.
- D=2 combinations did not converge to an improved solution before I left LANL.
  - Likely just needs more time.

- Not a home run result but proof of concept!
  - Other teams focus on:
    - predicting structure directly from spectra (go big or go home / challenging)
    - building better similarity metrics on just spectra (easy / limited improvements)
  - This shows that a hybrid approach is possible and may bridge the gap between existing methods and other approaches.

Side Project – Future Directions

- Niche enough that LANL let me take the code with me.
  - Open source on GitHub
- Need more performance for higher depth colors.
  - Mpi4py – go wide
  - Cython - make it faster
  - Calculations are all dot products / matrix math – maybe GPU?
- Individual models could be tuned further.
  - Automate hyperparameter searching per model
  - Could try different algorithms dynamically
    - SVM vs RF
  - 1-d convolutional techniques?
- Test on data from new instruments
  - NIST spectra are low resolution but new GC-MS systems have higher resolution.
  - Are the models resolution-dependent?

Publications

1. **RM Flight, Mitchell JM, Moseley HNB "Scan-Centric, Frequency-Based Method for Characterizing Peaks from Direct Injection Fourier Transform Mass Spectrometry Experiments" (2022) Metabolites**
2. **Mitchell JM, RM Flight, Moseley HNB "Untargeted Lipidomics of Non-Small Cell Lung Carcinoma Demonstrates Differentially Abundant Lipid Classes in Cancer vs. Non-Cancer Tissue" (2021) Metabolites**
3. Jin H, Mitchell JM, Moseley HNB "Atom Identifiers Generated by a Neighborhood-Specific Graph Coloring Method Enable Compound Harmonization across Metabolite Databases" (2020) Metabolites
4. Mitchell JM, Flight RM, Moseley HNB "Deriving Lipid Classification Based on Molecular Formulas" (2020) Metabolites
5. Mitchell JM, Flight RM, Moseley HNB "Small Molecule Isotope Resolved Formula Enumeration: A Methodology for Assigning Isotopologues and Metabolite Formulas in Fourier Transform Mass Spectra" (2019) Analytical Chemistry
6. Mitchell JM, Flight RM, Wang QJ, Higashi RM, Fan TWM, Lane AN, Moseley HNB "New Methods to Identify High-Peak Density Artifacts in Fourier Transform Mass Spectra and to Mitigate Their Effects on High-Throughput Metabolomic Data Analysis" (2018) Metabolomics
7. Trainor PJ, Mitchell JM, Clarlisle SM, Moseley HNB, DeFilippis AP, Rai SN "Inferring metabolite interactomes via molecular structure informed Bayesian graphical model selection with application to coronary artery disease" (2018) BioArxiv
8. Mitchell JM, Flight RM, Wang QJ, Higashi RM, Fan TWM, Lane AN, Moseley HNB "Development and *in silico* evaluation of large-scale metabolite identification methods using functional group detection for metabolomics" (2014) Frontiers in Genetics
9. Kannan S, Jones J, Mitchell JM "Dynamic Scheduling of White Water Rafting (2013) Harvard College Mathematics Review

Patents

1. Moseley HNB, Carrer WJ, Mitchell JM, Flight RM "Method and System for Identification of Metabolites using Mass Spectra" US Patent 10,607,723

Objectives for LANL Work

1. Submit 4 papers (datasets LA-UR'd so I can help after LANL)
   a. *In preparation.* Carter K., Mitchell JM*, et al. "The effect of Microbiome and Watering on Maize Root Exudates" (2022) New Phytologist?
   b. *In preparation.* Yeager C.M., Kaplan, D., Santschi, P., Mitchell JM, et al. "Iodide oxidation by forest soils is principally related to the activity of extracellular oxidase: (2022) Frontiers in Chemistry
   c. *In preparation.* Sevanto S., Mitchell JM, et al. *LDRD main experiment results*
   d. *In preparation.* Newman B., Mitchell JM, et al. *Wick Pot side experiment results*
   e. Machine Learning Work? Depends on if it works or not.
2. Further develop collaborative relationships
   - UKY – Will be finishing two software projects from MD/PhD with Hunter Moseley
3. Continue to develop broad technical skills necessary for transition to industry
   - Rounding out my Comp Sci. skillset for transition to Med-Tech (Dev Ops / Software Packaging)
   - Take my patent and make something I can sell to Thermo (needs GUI, documentation, packaging)

Acknowledgements

LANL

- Chris Yeager (Mentor, C-CDE)
- Nicholas Lubbers (Co-Mentor, CCS-3)
- Sanna Sevanto (Co-Mentor, EES-14)
- Kelsey Carter (EES-14)
- Eric Moore (B-IOME)
- Brent Newmann (EES-14)
- Tom Yoshida (C-CDE)
- Dan Huber (C-CDE)
- Phillip Mach (B-11)
- Anastasiia Kim (CCS-3)
- Chris Freye (Q-5)

UKY (helped with stats)

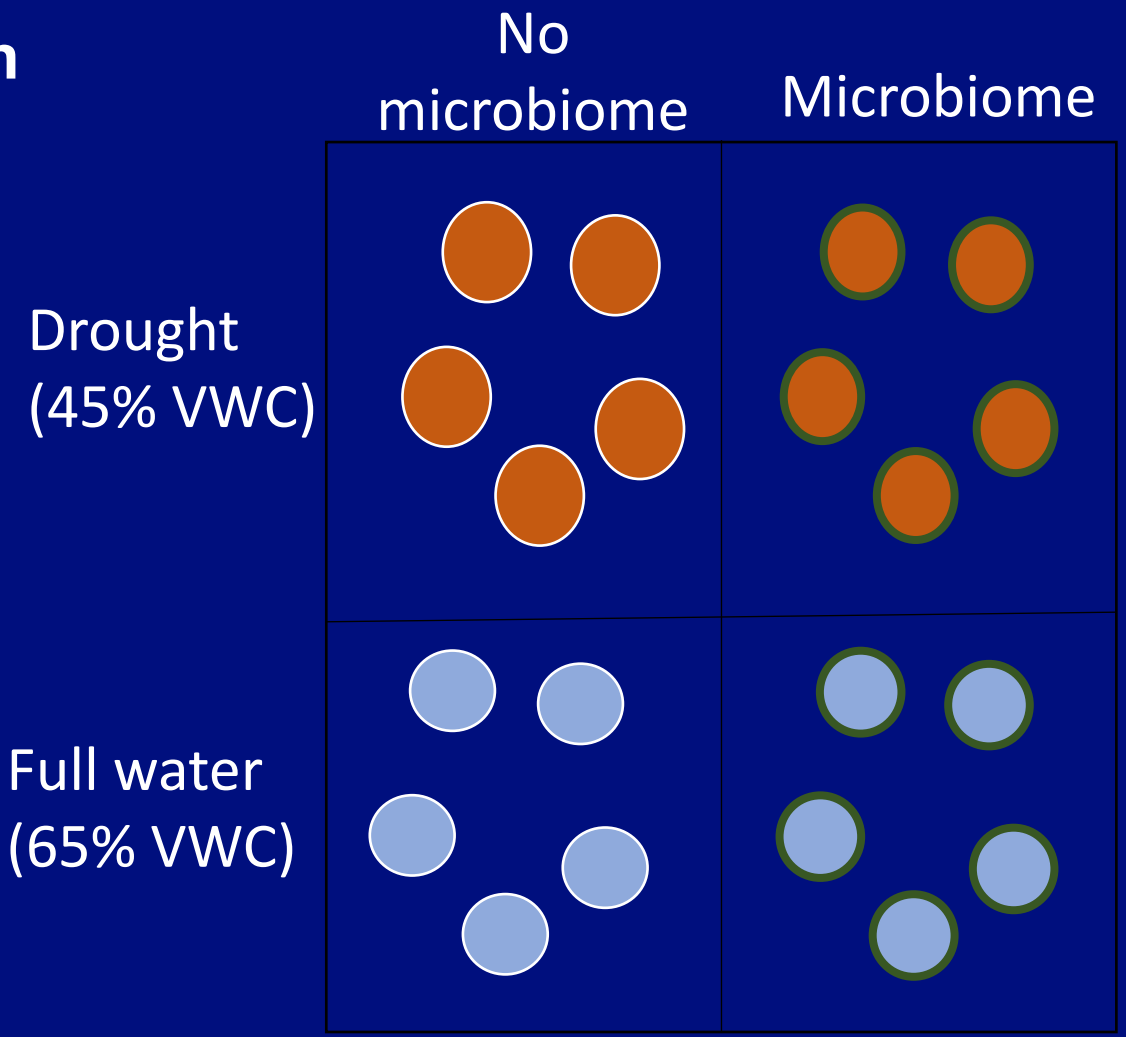
- Robert Flight
- Hunter Moseley

U. Wash (helped with GCxGC-TOFMS)

- Rob Synovec

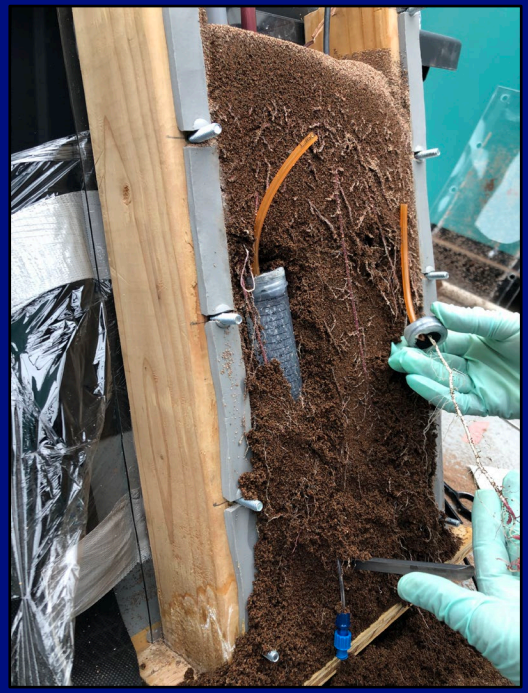
This research used resources provided by the Darwin testbed at Los Alamos National Laboratory (LANL) which is funded by the Computational Systems and Software Environments subprogram of LANL's Advanced Simulation and Computing program (NNSA/DOE).

Design

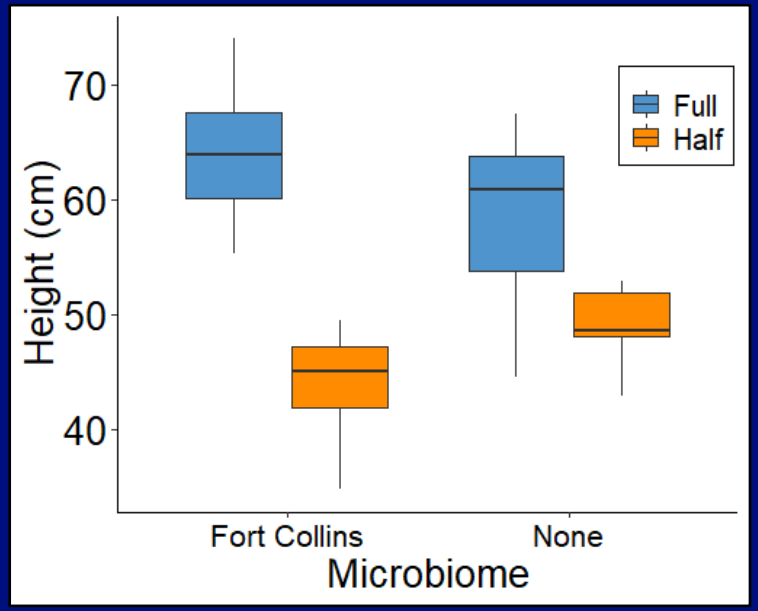


Conducted and planned analyses

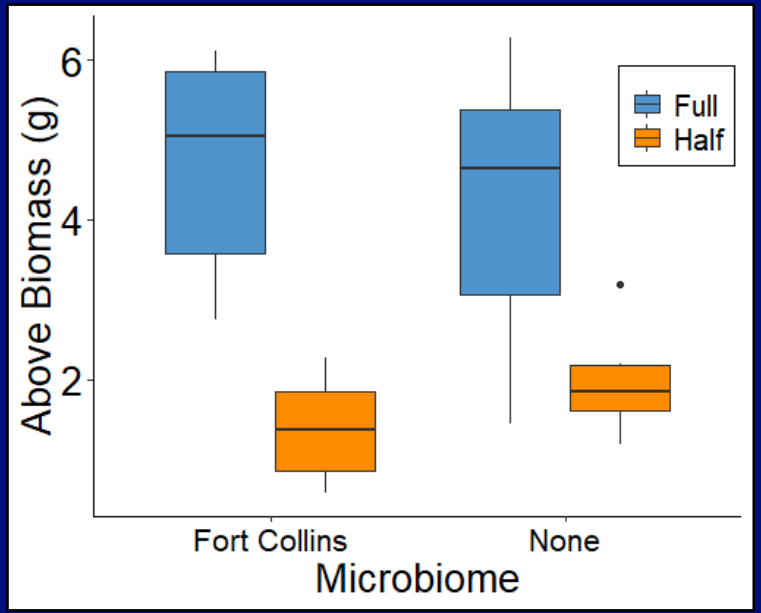
- Measured plant traits – Above and belowground growth
  - Photosynthesis and stomatal conductance
  - Leaf traits (area, %water content, LMA)
- Carbohydrate and invertase analyses – roots and exudates
- Leaf, root, exudate metabolomics
  - GC-MS, LC-MS



Water and microbiome treatment – stronger effect on aboveground growth



- With microbiome – full water plants taller than half water
- No microbiome – no difference between water treatments



- Aboveground dry mass lower in half water

Root Metabolic Profiles Change in Response to Microbiome and Watering Treatments

Differentially Abundant Metabolites (Microbiome, p<.05)

Carbohydrates:
- 1,5-Anhydroglucitol
- 3-a-Mannobiose
- **D-Turanose**
- **Sucrose**
- Maltose

Organic Acids:
- Quininic Acid
- Ribonic Acid
- Mannonic Acid
- Galactaric Acid
- Citric Acid

Butanedioic Acid*

Fatty Acid:
- Palmitic Acid*

← More Abundant in No Microbiome

← More Abundant in Fort Collins

Differentially Abundant Metabolites (Watering, p<.05)

Carbohydrates:
- **D-Turanose**
- **Sucrose**

Organic Acids:
- Shikimic Acid
- 4-Coumaric Acid

Phenolic/Polyol:
- Arbutin

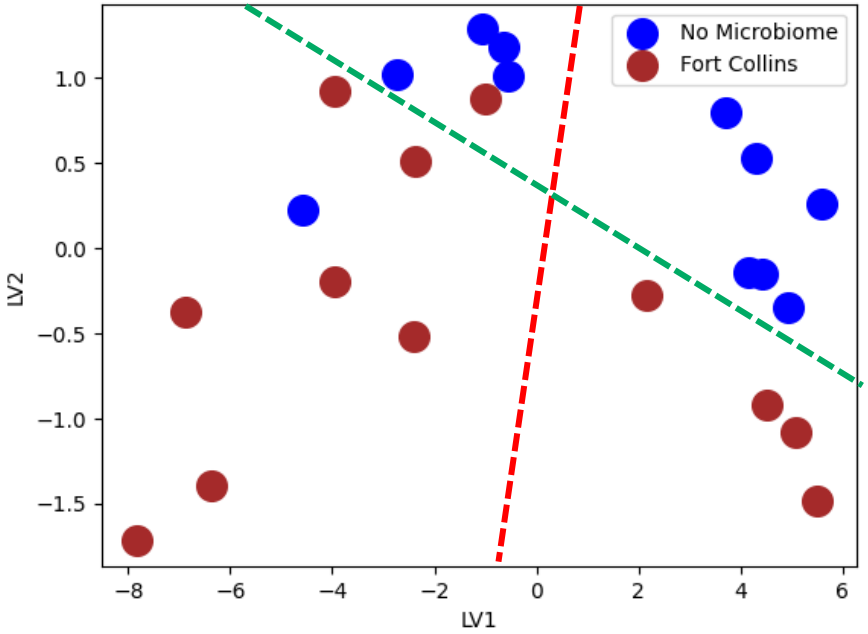
← More Abundant in Full Water

Fatty Acid:
- Nonanoic Acid*

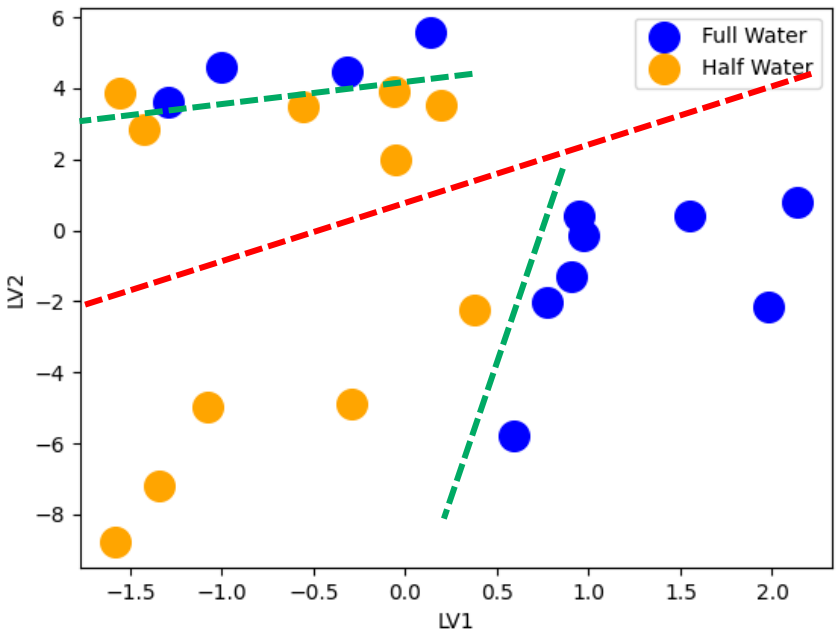
← More Abundant in Half Water

Unidentified but Differentially Abundant Metabolites Omitted!

PLSDA models trained on features selected in an unbiased manner



PLSDA can differentiate microbiome treatments



PLSDA can partially differentiate watering treatments

Work done in conjunction with Dr. Lubbers and Dr. Kim

GCxGC-TOFMS Chromatograms of Pooled Root and Exudate Samples

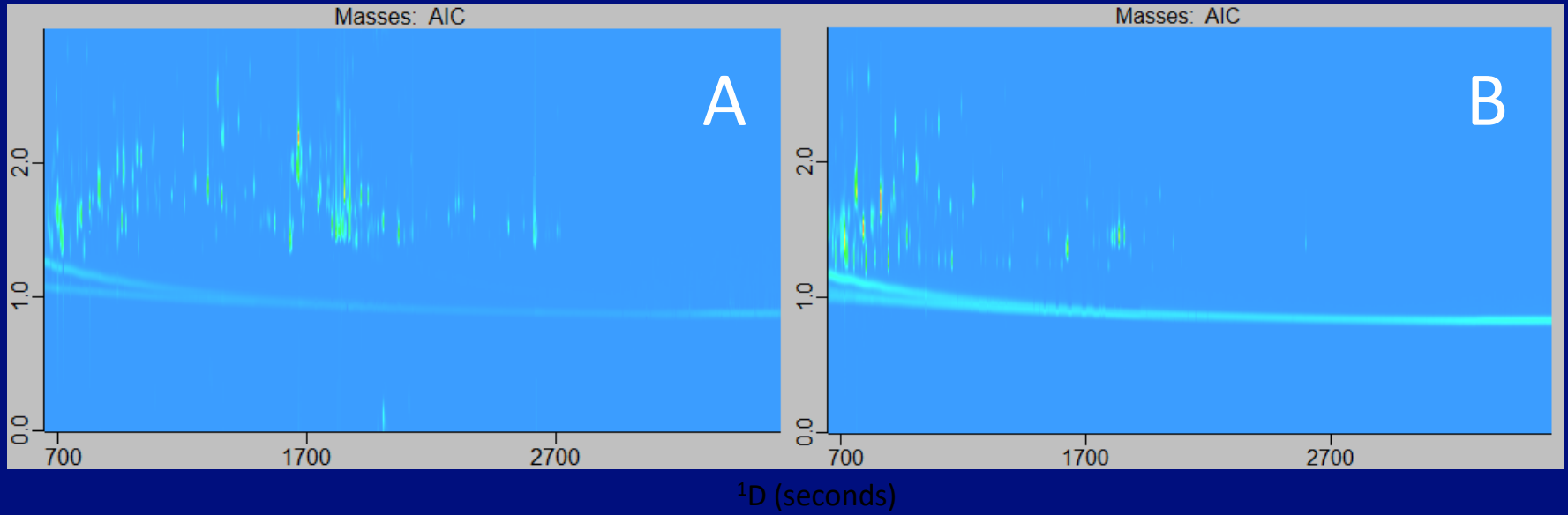


Figure ___: Two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) chromatograms of pooled root and exudate samples (panels A and B respectively). ^{1}D and ^{2}D represent the retention times of a given compounds on the primary and secondary columns, a 30m Rxi-5Sil and a 1.0m DB-17MS respectively. The distribution of peaks with similar ^{1}D times along the ^{2}D highlights the ability of two-dimensional chromatography to separate better complex samples commonly observed in metabolomics studies. Unsurprisingly, our root tissue samples are more chemically complex and clearly more concentrated than our exudate samples.

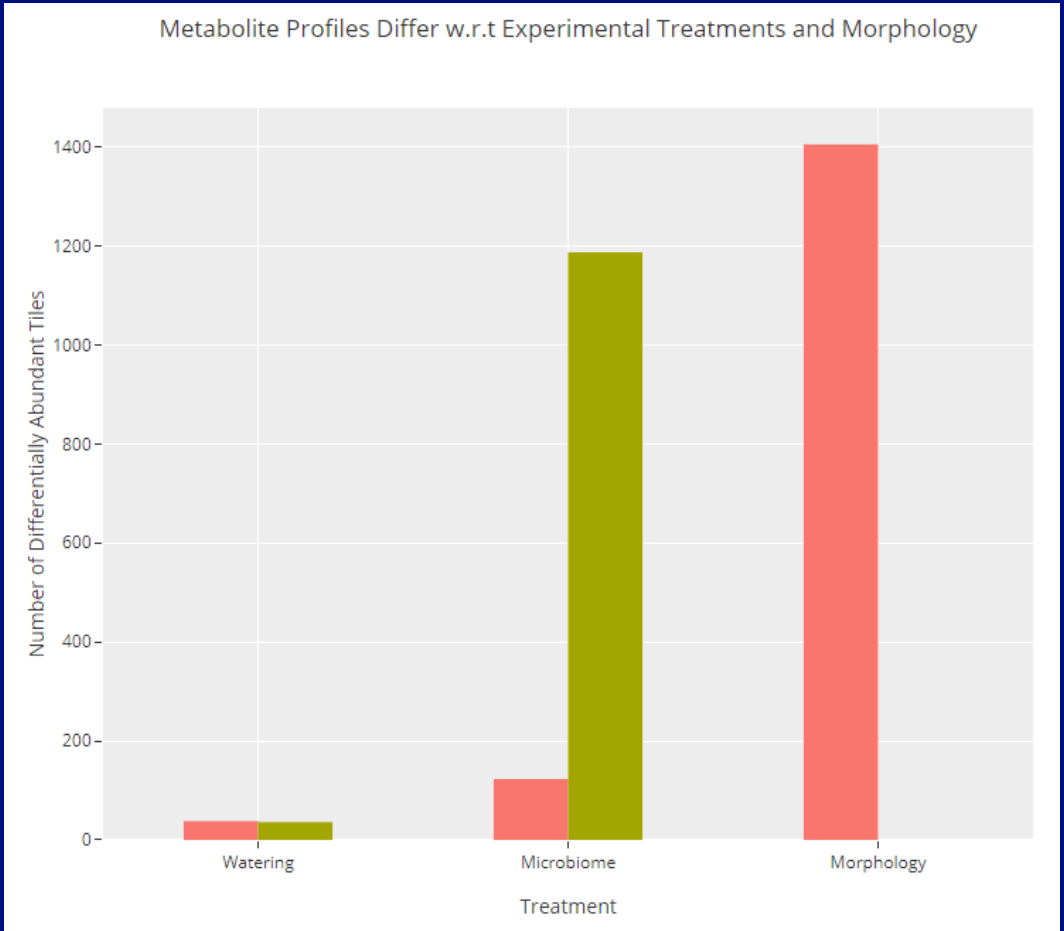
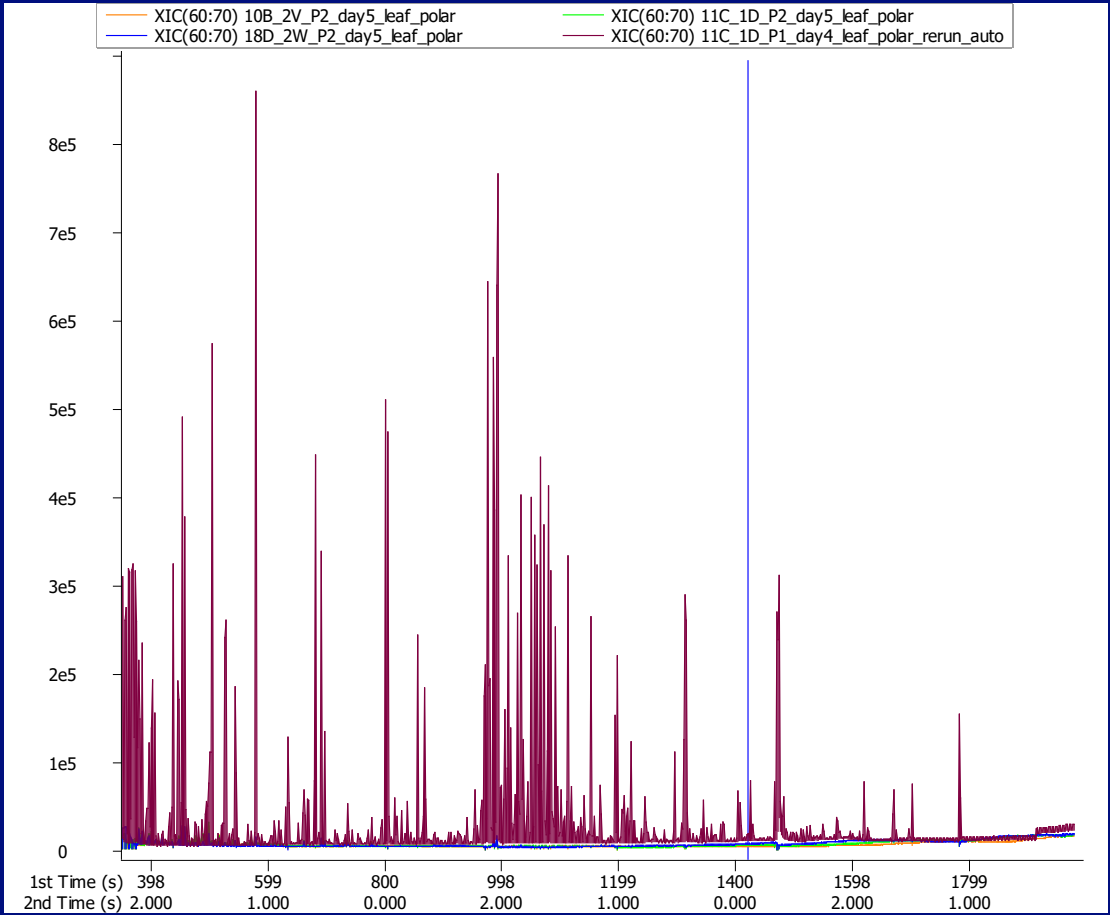
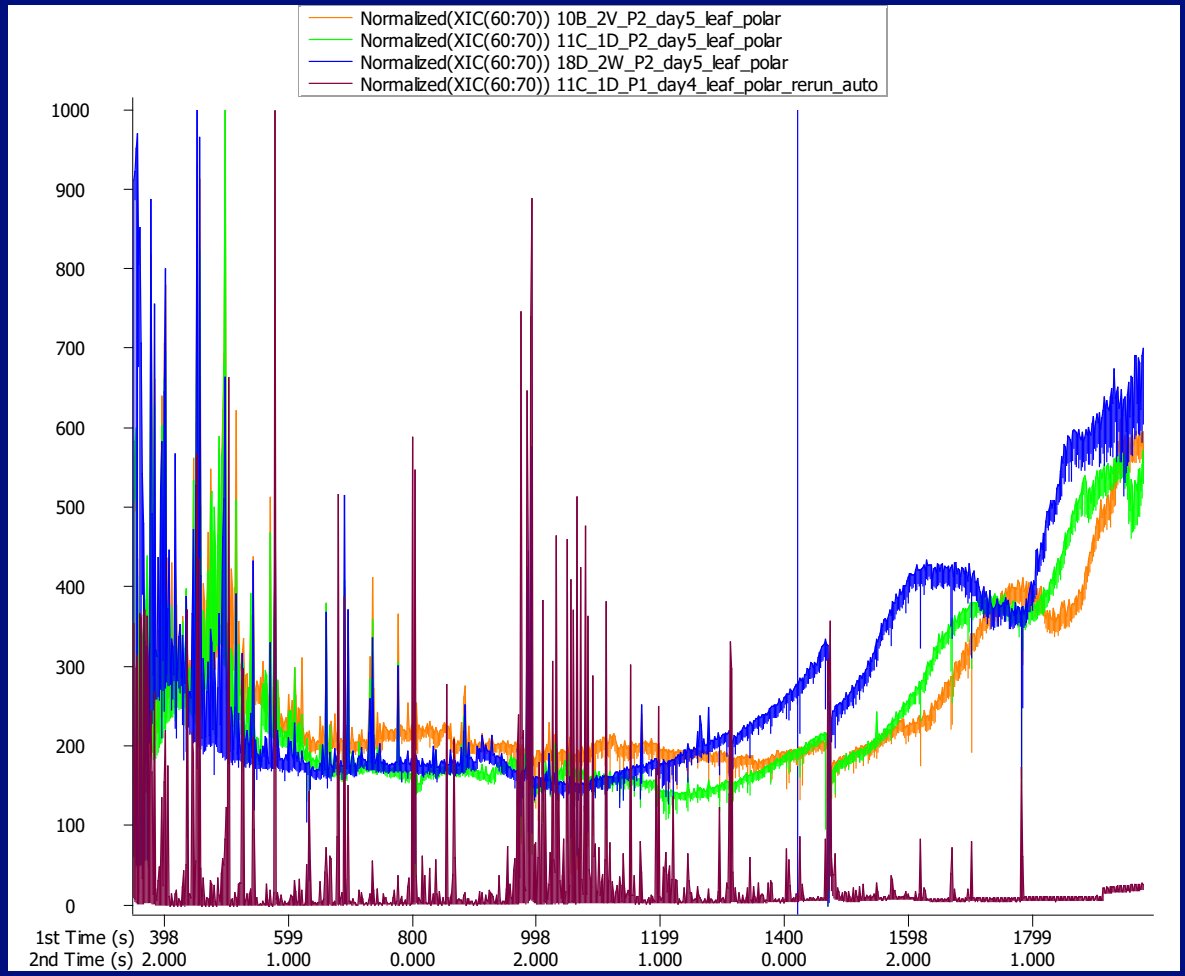
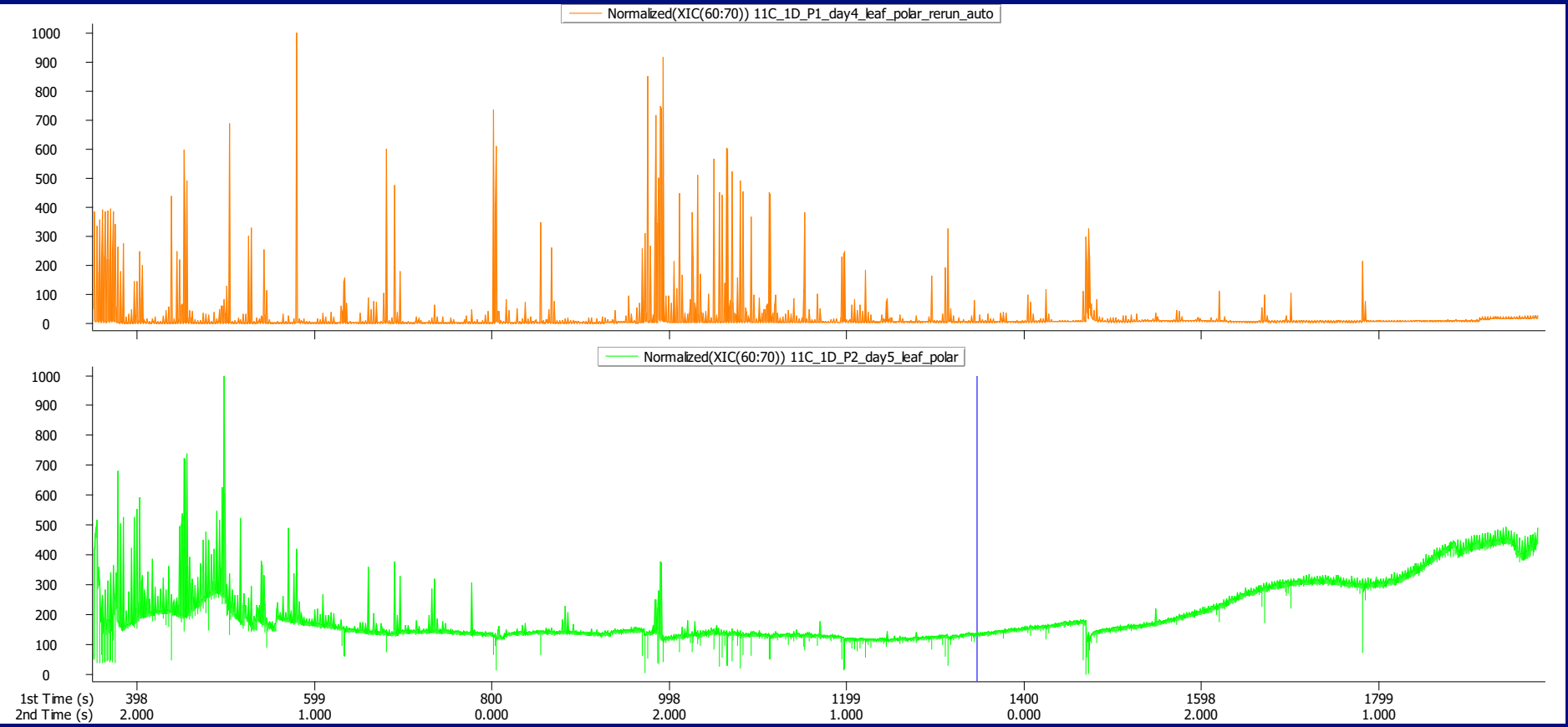


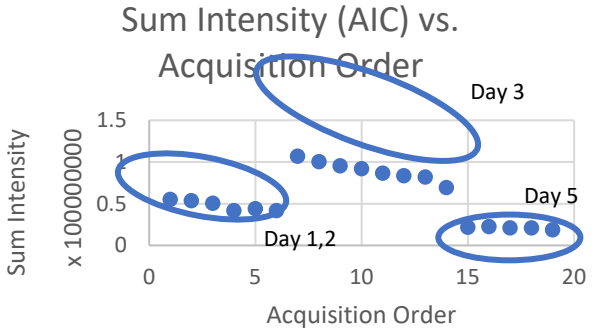
Figure ___: Metabolite profiles of roots (red) and exudates (green) differ with respect to watering and microbiome treatments. Metabolite profile differences were detected using ChromaTof Tile that compares chromatograms by dividing them into tiles and performing a Fischer-ratio based test on the mass spectrum for each tile across treatments. The number of tile hits ($p<0.05$), partially reflects the metabolite differences between treatments. While not a true treatment, dividing the roots into two morphological groups based on median root diameter demonstrates that although watering and microbiome treatments induce some metabolic differences, they are small compared to the effect of morphology in roots. The absolute number of hits is greater in the exudate than expected; however, false positive hits from the presence of fertilizer compounds exaggerates the number of hits in exudate samples.

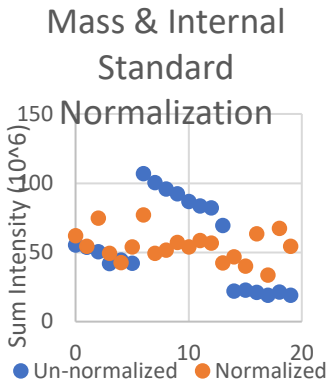

XIC(60:70) 10B_2V_P2_day5_leaf_polar
XIC(60:70) 18D_2W_P2_day5_leaf_polar
XIC(60:70) 11C_1D_P2_day5_leaf_polar
XIC(60:70) 11C_1D_P1_day4_leaf_polar_rerun_auto
8e5
7e5
6e5
5e5
4e5
3e5
2e5
1e5
0
1st Time (s) 398 599 800 998 1199 1400 1598 1799
2nd Time (s) 2.000 1.000 0.000 2.000 1.000 0.000 2.000 1.000

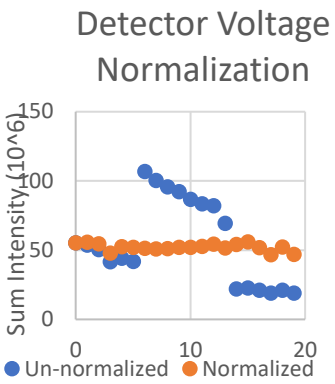

Normalized(XIC(60:70)) 10B_2V_P2_day5_leaf_polar
Normalized(XIC(60:70)) 11C_1D_P2_day5_leaf_polar
Normalized(XIC(60:70)) 18D_2W_P2_day5_leaf_polar
Normalized(XIC(60:70)) 11C_1D_P1_day4_leaf_polar_rerun_auto
1000
900
800
700
600
500
400
300
200
100
0
1st Time (s) 398 599 800 998 1199 1400 1598 1799
2nd Time (s) 2.000 1.000 0.000 2.000 1.000 0.000 2.000 1.000

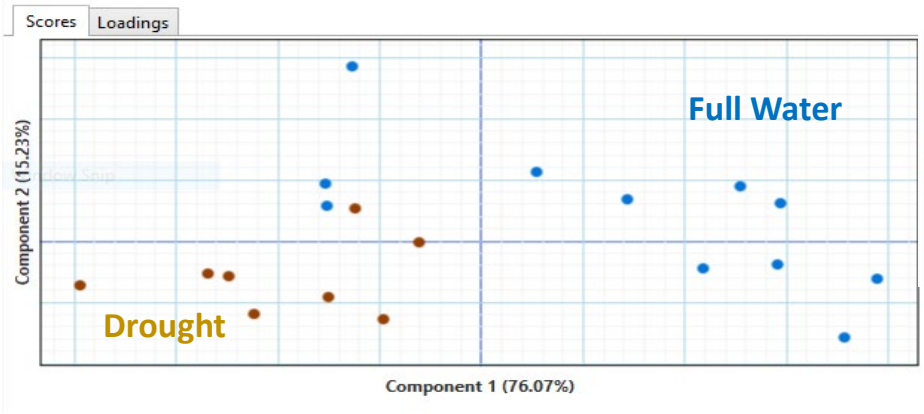

Normalized(XIC(60:70)) 11C_1D_P1_day4_leaf_polar_rerun_auto
1000
900
800
700
600
500
400
300
200
100
0
Normalized(XIC(60:70)) 11C_1D_P2_day5_leaf_polar
1000
900
800
700
600
500
400
300
200
100
0
1st Time (s) 398 599 800 998 1199 1400 1598 1799
2nd Time (s) 2.000 1.000 0.000 2.000 1.000 0.000 2.000 1.000

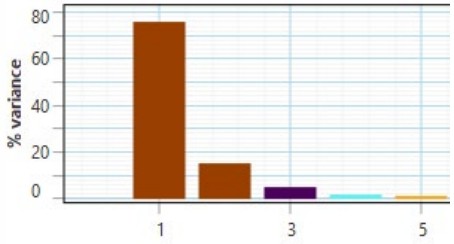
Data Normalization and PCA (Leaf Tissue)







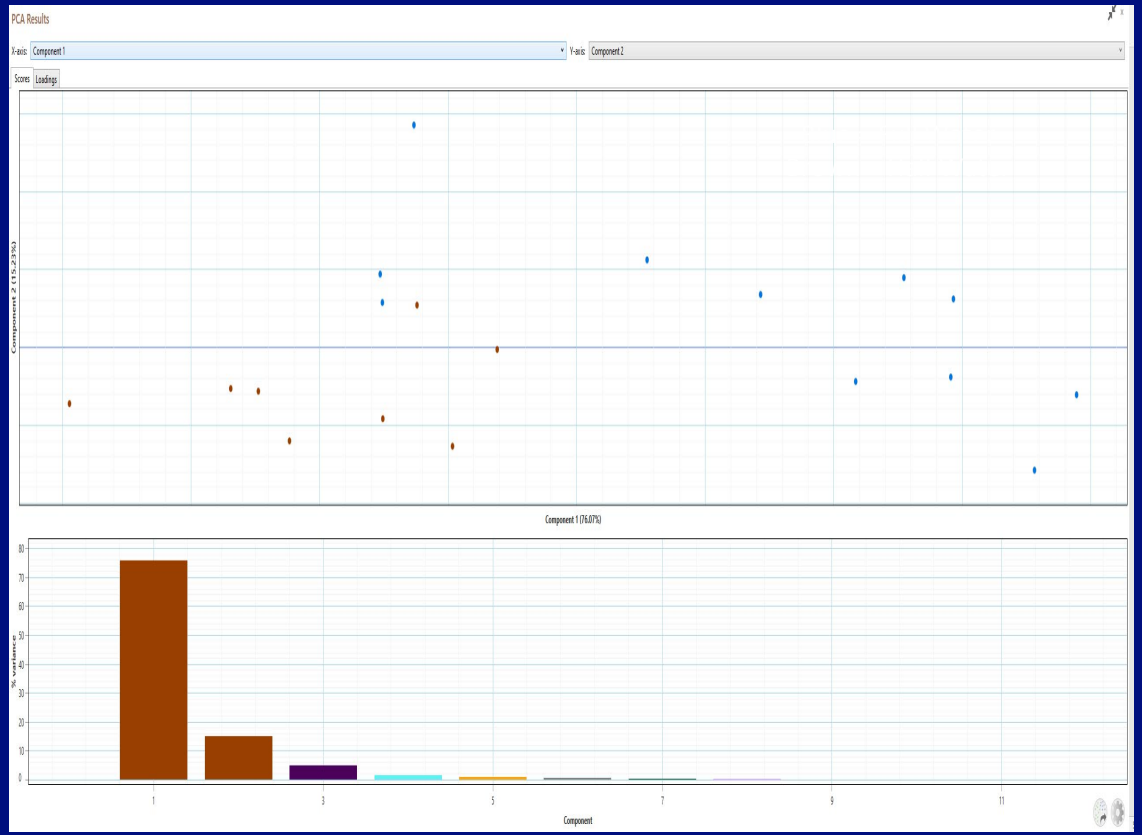


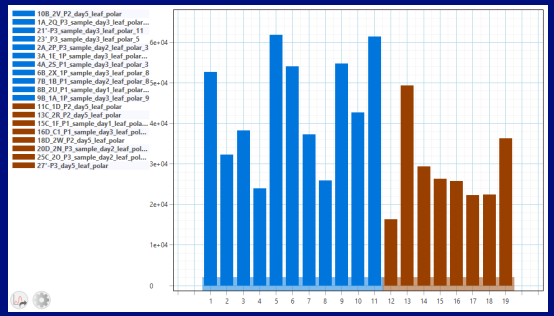


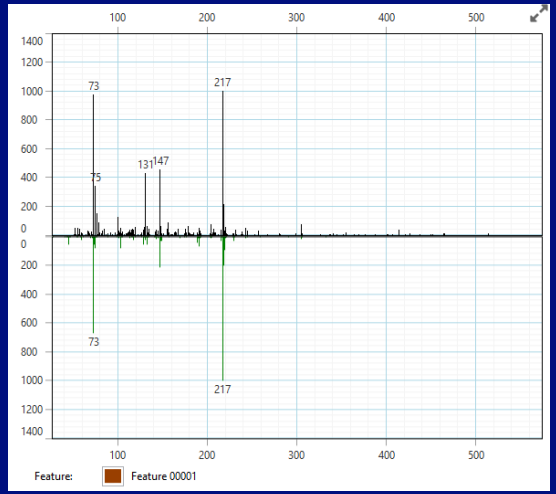
Just 3 Components Capture >85% of the Variance!

Metabolite Profiles Separate in PCA Space

Normalized by Mass and Ribitol (2d-Integral)
Chromatof Tile Results – Water Status







Normalized by Mass and Ribitol (2d-Integral) Chromatof Tile Results

Contaminants Or Misassignments

Loadings

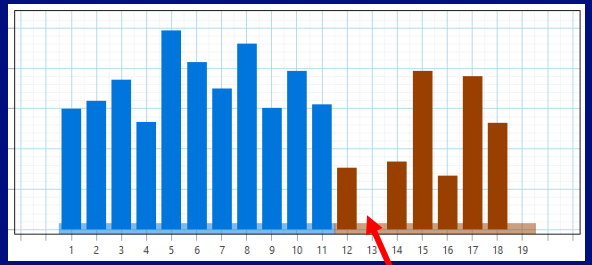
Feature 00068
- PC1: .85
- PC2: -.38

Methoxyamine 2TMS:
- PC1: .28
- PC2: .77

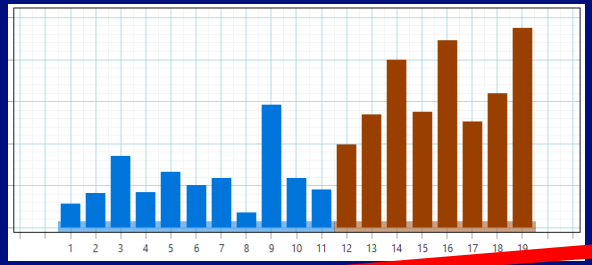
Id		Name	Formula	Similarity	Reverse	Probability (%)	CAS
1 (1)		Feature 00001		0	0	0.00	
67 (2)		Feature 00067		0	0	0.00	
68 (2)		Feature 00068		0	0	0.00	
82 (3)		Feature 00082		0	0	0.00	
144 (4)		Feature 00144		0	0	0.00	
183 (5)		Cyclotrisiloxane, hexamethyl-	$C_6H_{18}O_3Si_3$	851	851	37.07	541-05-9
246 (6)		Feature 00246		0	0	0.00	
285 (7)		Feature 00285		0	0	0.00	
347 (8)		Methoxyamine, 2TMS derivative	$C_7H_{21}NOSi_2$	896	896	96.84	7266-76-4
454 (9)		Lactulose, octakis(trimethylsilyl) etl	$C_{37}H_{89}NO_{11}Si_8$	783	783	14.02	
480 (10)		Feature 00480		0	0	0.00	
522 (11)		Feature 00522		0	0	0.00	
530 (12)		Aucubin, hexakis(trimethylsilyl) eth	$C_{33}H_{70}O_9Si_6$	801	801	11.25	
586 (13)		Feature 00586		0	0	0.00	
601 (14)		Methoxyamine, 2TMS derivative	$C_7H_{21}NOSi_2$	857	857	92.55	7266-76-4
738 (15)		Quininic acid (5TMS)	$C_{22}H_{32}O_6Si_5$	791	791	76.09	
918 (16)		Trisiloxane, 1,1,1,5,5,5-hexamethyl-	$C_{12}H_{36}O_4Si_5$	881	881	78.83	3555-47-3
961 (16)		Trisiloxane, 1,1,1,5,5,5-hexamethyl-	$C_{12}H_{36}O_4Si_5$	888	888	79.20	3555-47-3
988 (17)		D-Xylofuranose, 1,2,3,5-tetrakis-O-(	$C_{17}H_{42}O_5Si_4$	824	824	25.73	56271-68-2
1079 (18)		Feature 01079		0	0	0.00	
1124 (19)		Feature 01124		0	0	0.00	
1174 (20)		D-(+)-Galactose oxime (isomer 1), (	$C_{24}H_{61}NO_6Si_6$	863	863	5.01	

Unknown Compounds

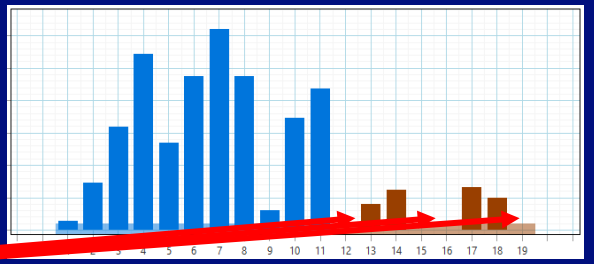
Feature 00522



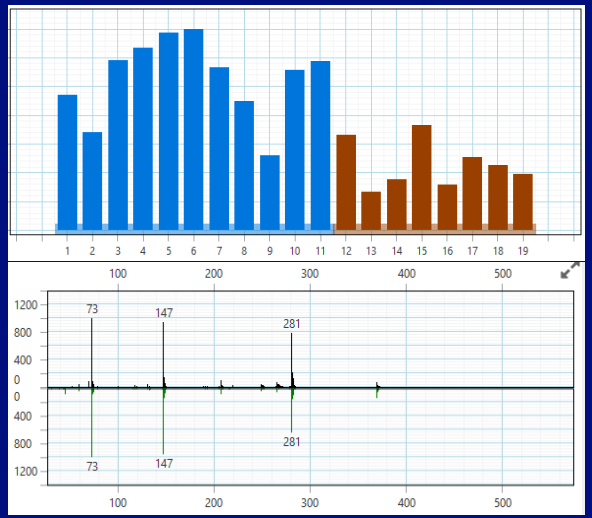
Feature 00246



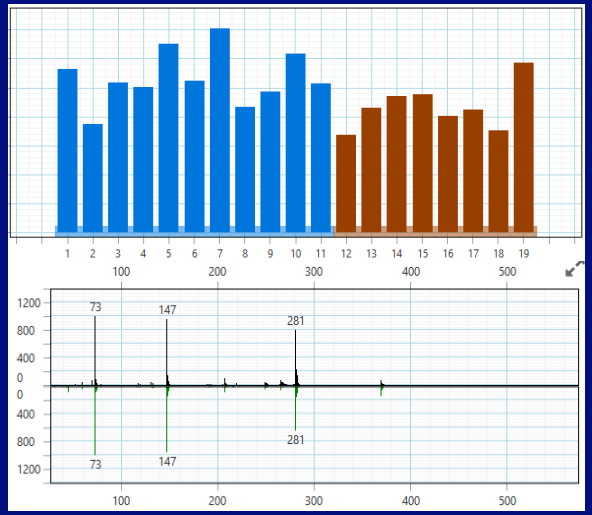
Feature 01443



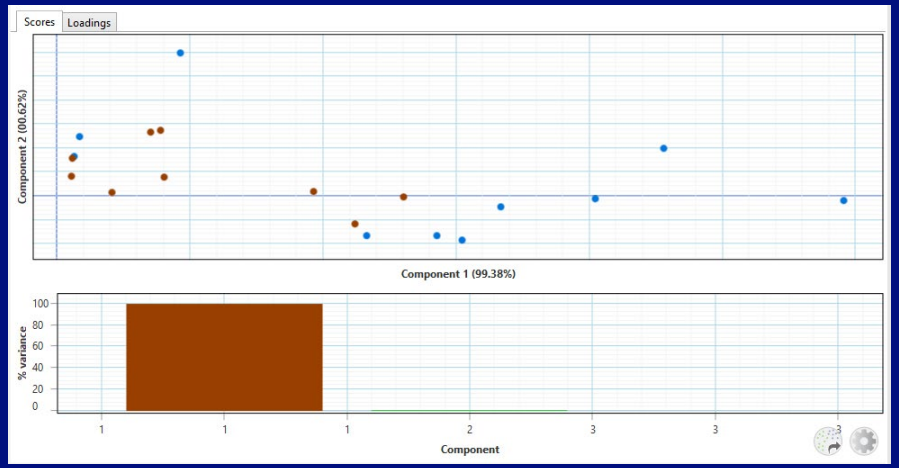
Missing Value = 0?



Trisiloxane

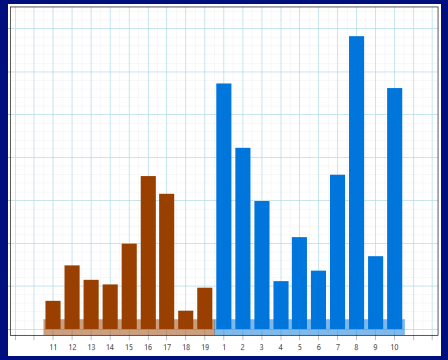


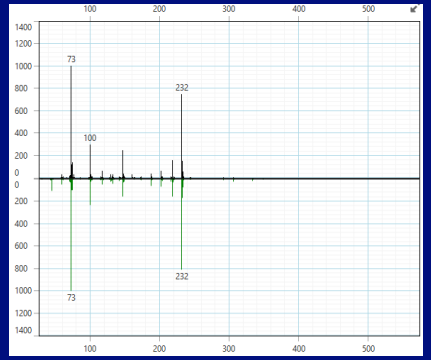
Trisiloxane



Normalized by Mass and Ribitol (2d-Integral)
Chromatof Tile Results
Microbiome (Brown) or No Microbiome (Blue)

Id		Name	Formula
1 (1)		L-Aspartic acid, 3TMS derivative	$C_{13}H_{31}NO_4Si_3$
5 (1)		L-Aspartic acid, 3TMS derivative	$C_{13}H_{31}NO_4Si_3$
10 (2)		Feature 00010	
13 (3)		Feature 00013	
17 (4)		Feature 00017	
23 (5)		Feature 00023	

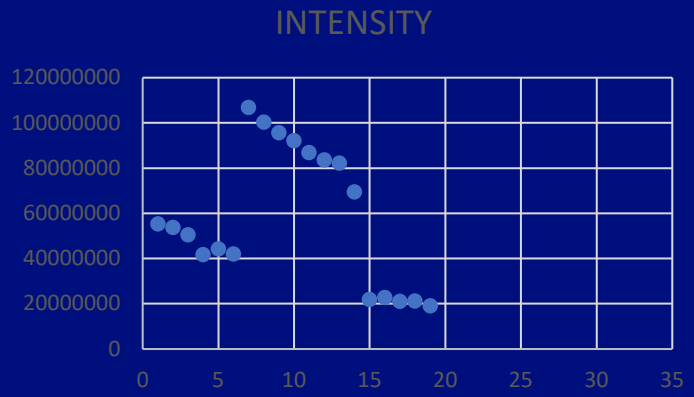


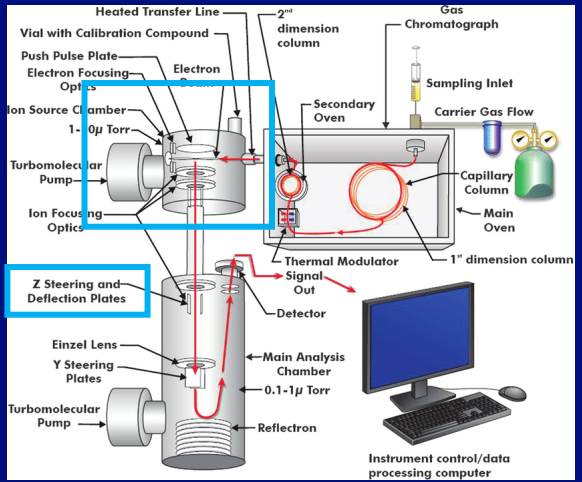


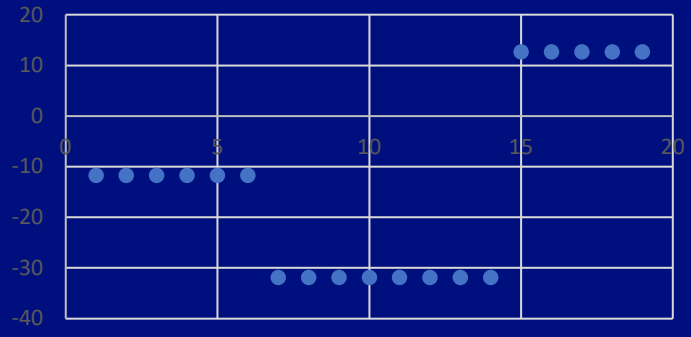
Aspartic Acid

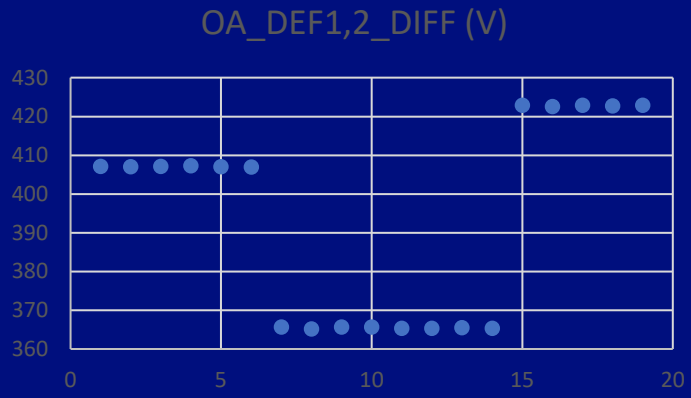
90.78%
similarity

Voltage Corrections?



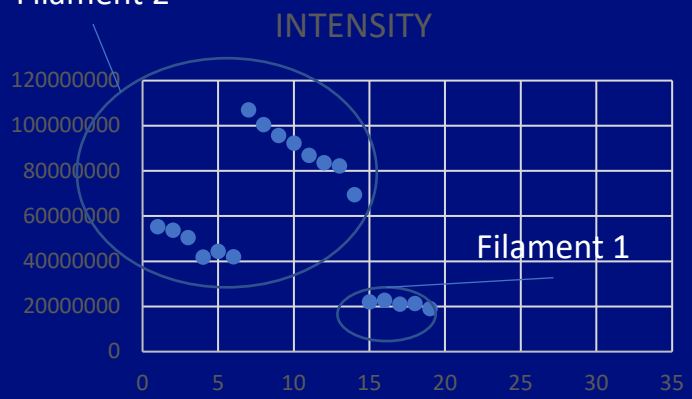


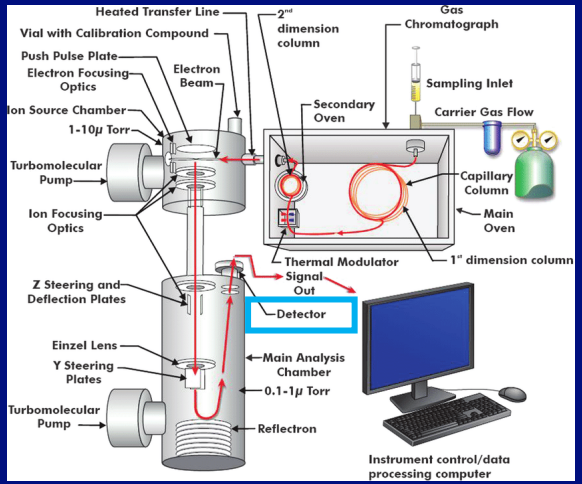


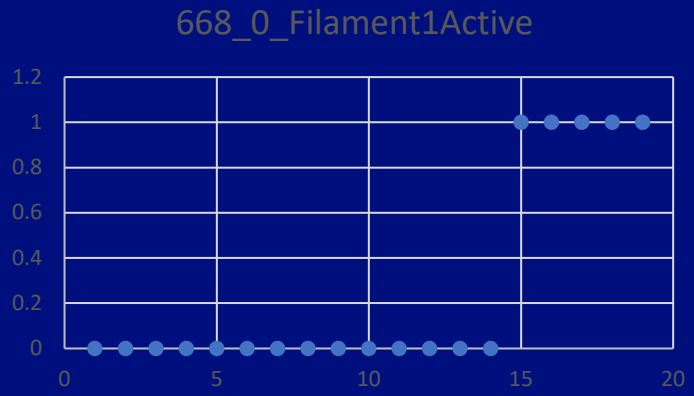


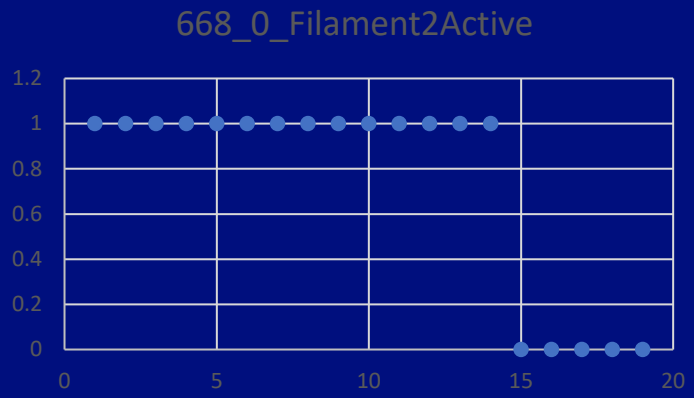
Filament Corrections?

Filament 2

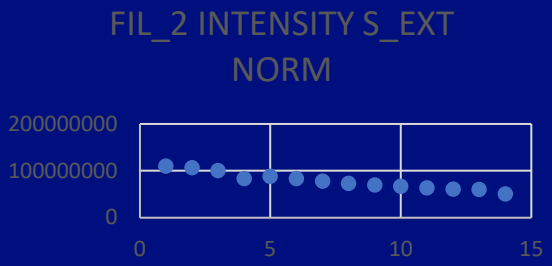


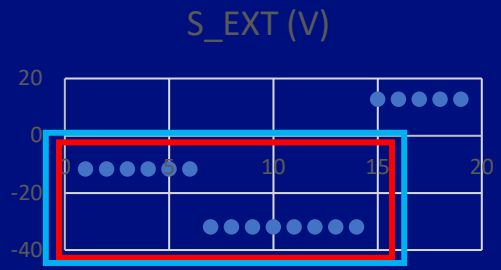




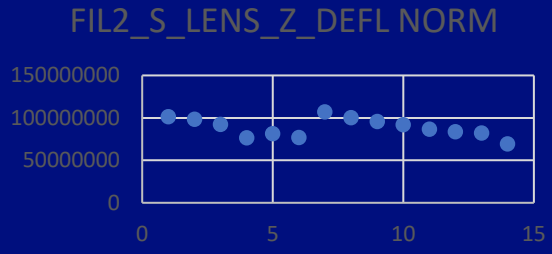


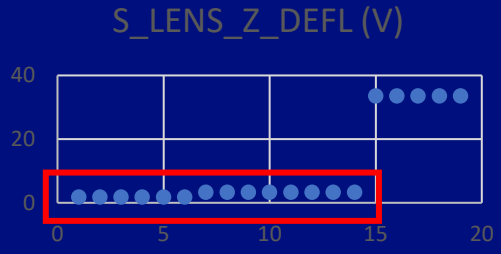
Not sure if S_EXT is the voltage for the Source or the S_LENS

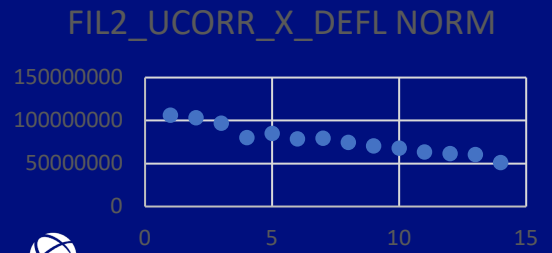


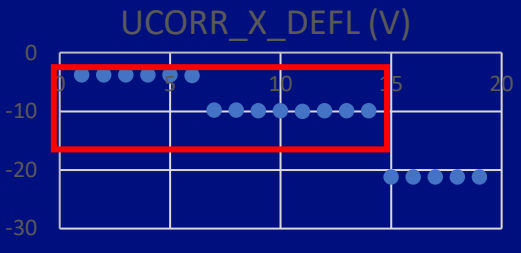


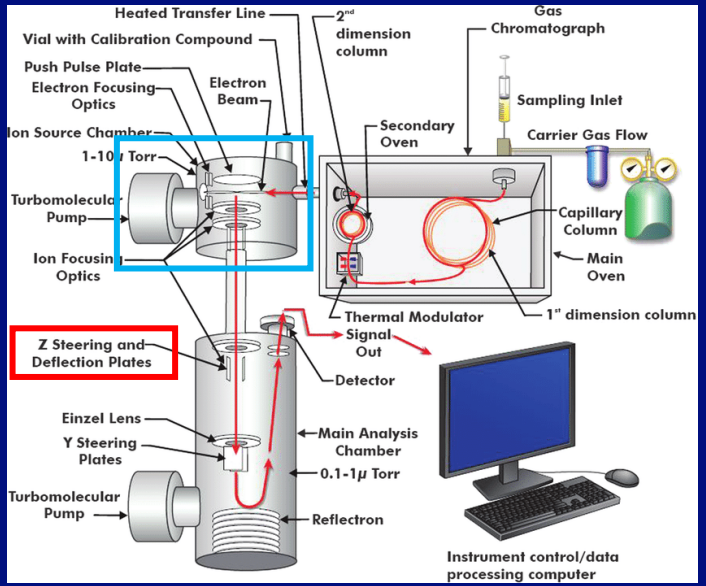
The S_LENS focuses ions into a beam to move through the instrument

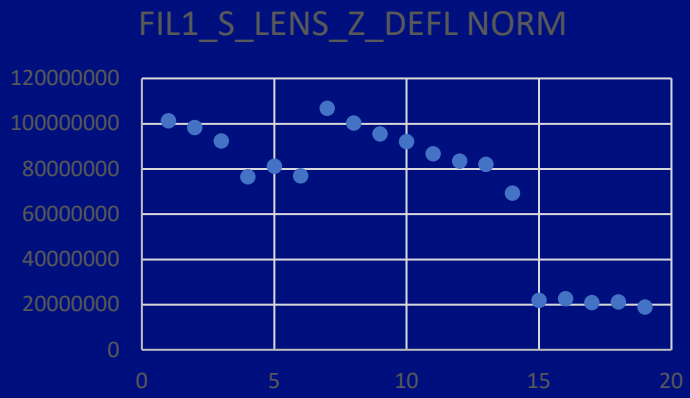


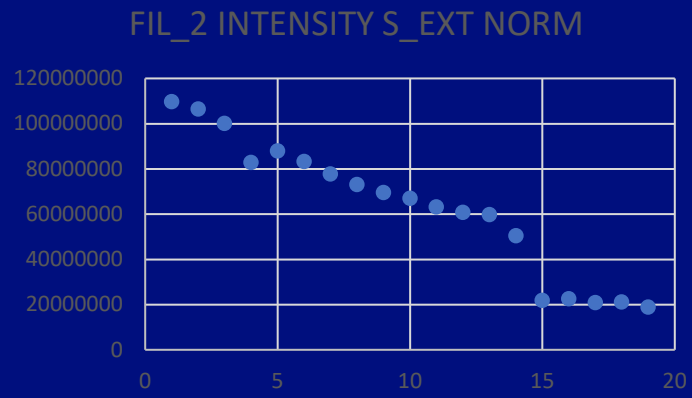




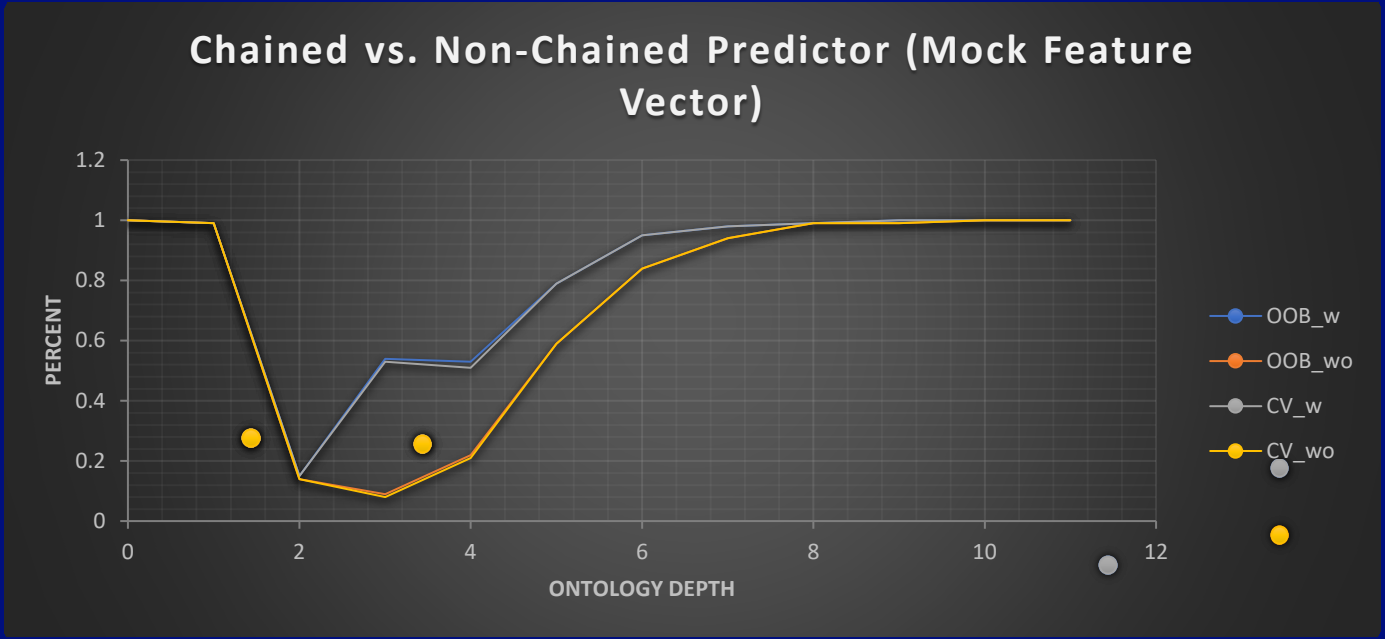


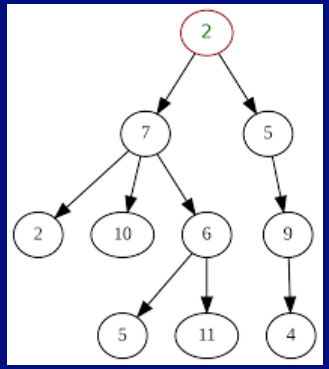






What about Filament 2?



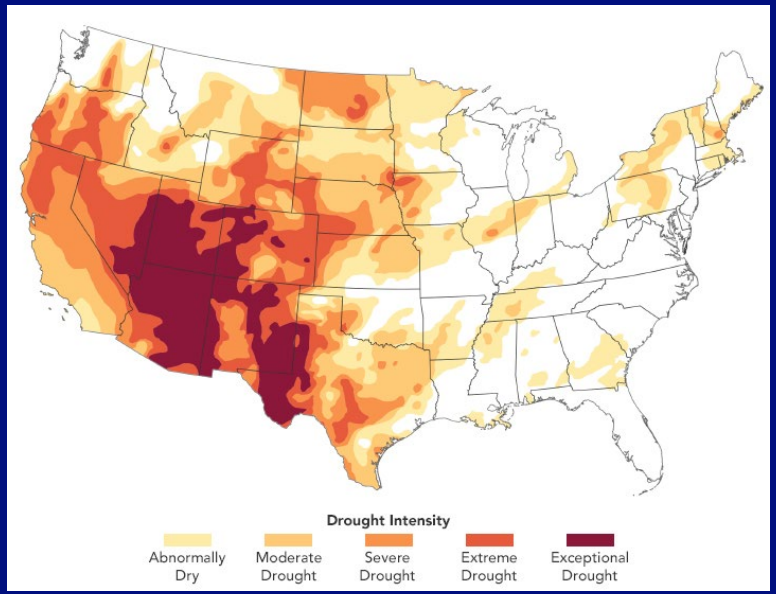


Chained = Add predicted parent class to feature vector at each subsequent depth.

- Examples are very sparse at low depth.
- Training dataset is all metabolites, introduces biases:
  - if you know the compound is an Amino Acid, it is almost certainly an L-Amino Acid.
  - If it's a sugar, it's a D-Sugar.

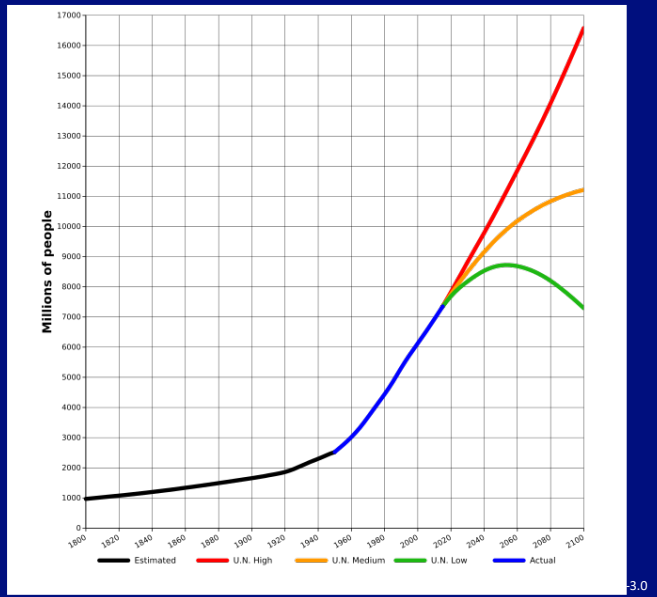
Need More Examples

Climate Change and Population Growth Threaten Food and Biofuel Security



Credit: nasa.gov

Climate Change = Decreased Crop Yield



Population Growth = Increased Crop Demand

Can we modify crops, soil or both to maintain sufficient food and biofuel production?

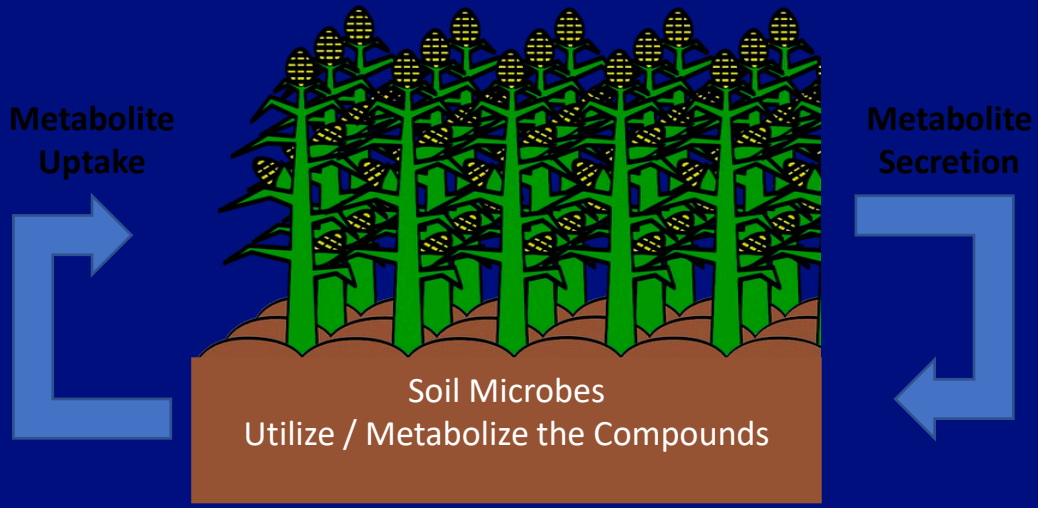
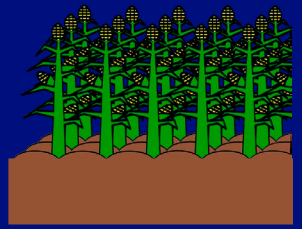
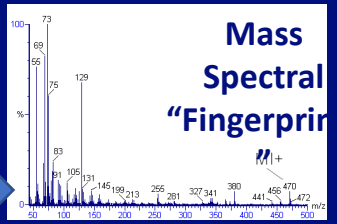


Image by Ehecoatl (pixabay), pixabay license

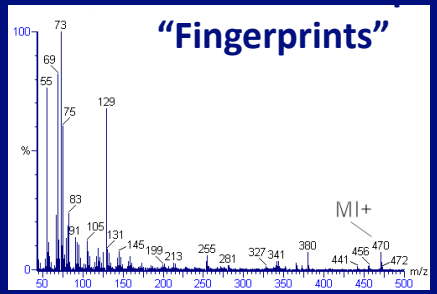


1
Sample

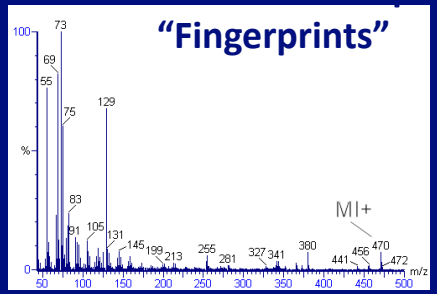
Credit: leco.com



"Mass_spectrum_brassicasterol.png" by Smmudge, CC-BY-SA-3.0



>50% Are Unidentifiable!



Metabolite Databases

>50% Are Unidentifiable!

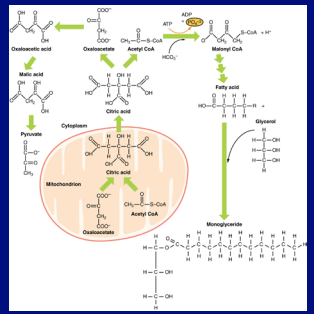
Machine Learning Models

Graph Theory

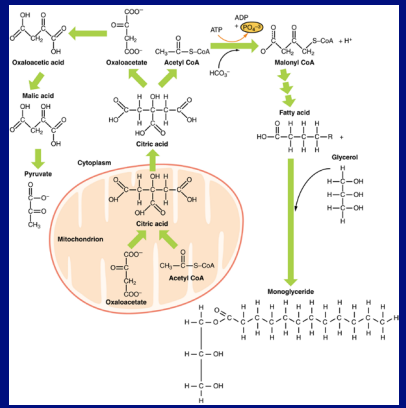
Advanced Analytical Chemistry

Comprehensive Metabolite Identification Without Databases!

Improved Metabolic Models



"_lipid_metabolism.png" by OpenStax College, CC-BY-SA-3.0


Oxaloacetic acid
Oxaloacetate
Acetyl CoA
ATP
ADP
Malonyl CoA
Malic acid
Citric acid
Fatty acid
Pyruvate
Cytoplasm
Glycerol
Citric acid
Mitochondrion
Acetyl CoA
Oxaloacetate
Monoglyceride