

Improving the Robustness and Toxicological Significance of Nontarget Chemical Identification in High Resolution Mass Spectrometric Data

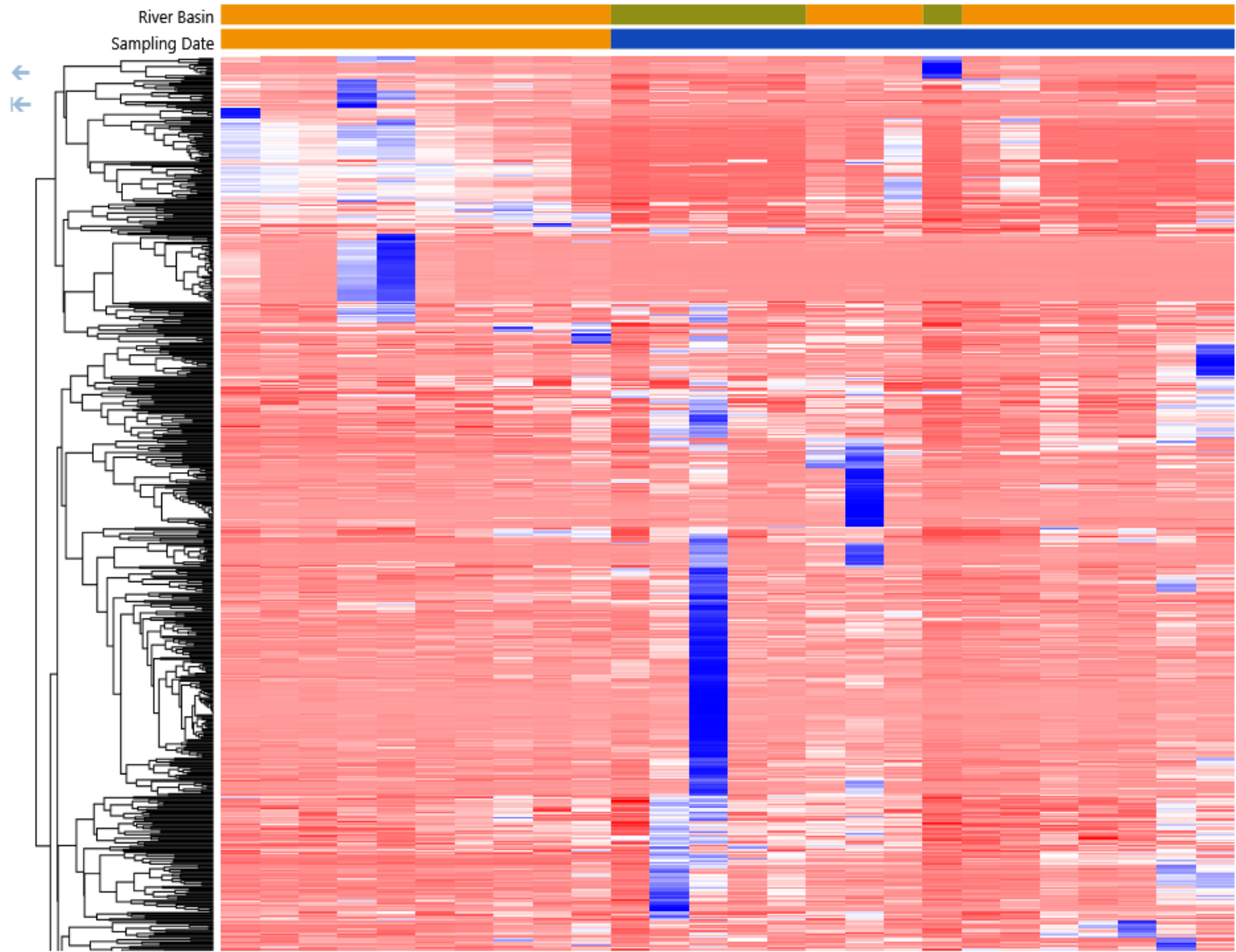
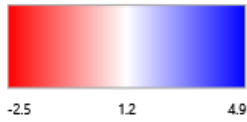
- Research Question: How can disparate sources of high-resolution MS data be combined to harmonize approaches for non-targeted environmental analysis

Collaborating SRP-Funded Grantees:

- **Duke University**: P. Lee Ferguson Ph.D. (PI, DUSRC Analytical Chemistry Core), Gordon J. Getzinger, Ph.D. (PI, DUSRC Data Supplement IUC)
- **University of California, Davis**: Thomas M. Young, Ph.D. (PI, UCD SRC Project 1: Bioremediation), Ilias Tagkopoulos Ph.D. (Bioinformatics and Data Science)
- **US EPA**: Antony Williams, Ph.D. (National Center for Computational Toxicology): external collaborator



Data Sets for Cross Validation: Duke



Hurricane Florence Water Pollution

- Non-targeted analysis results from ESI(+) HRAM MS/MS analysis of North Carolina river water
- Samples taken during and after major flooding from hurricane
- 2,337 Compounds detected
- 641 Compounds with spectral library match > 75%
- Compound abundance profiles were highly sample-dependent.

River Basin
■ Neuse
■ Cape Fear

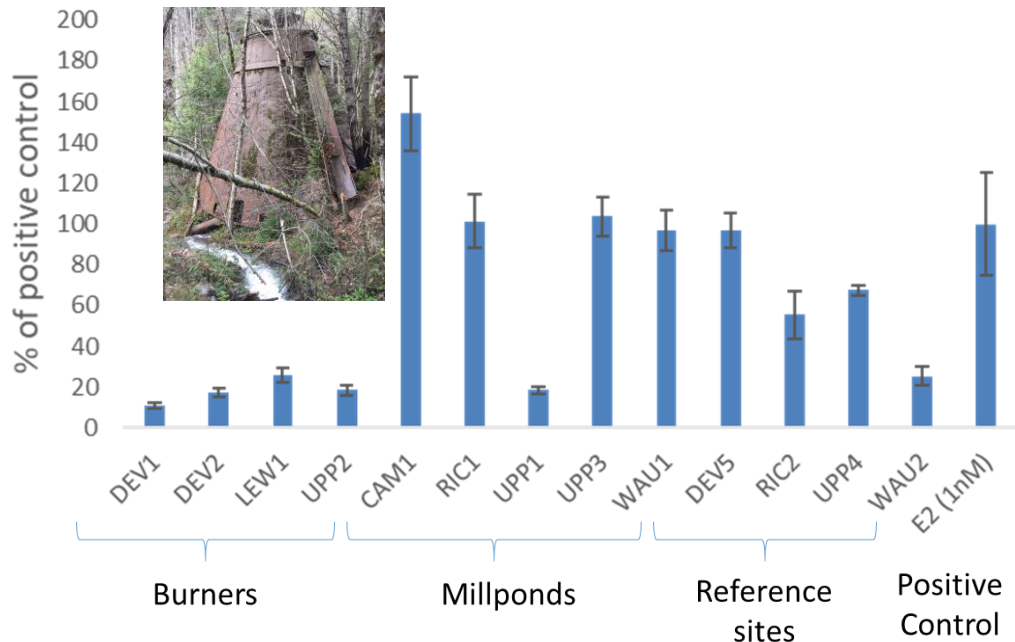
Sampling Date
■ 9/19/18
■ 10/10/18

Data Sets for Cross Validation: UC Davis



Identity of Compounds Correlated with Endocrine Activity Sought Using HRAM MS/MS

Abandoned Lumber Processing Sites on Yurok Tribal Lands Show Elevated Endocrine Activity



Edit this slide

F indable

A ccessable

I nteroperable

R eusable

What is FAIR?

Inputs: Challenges and Opportunities

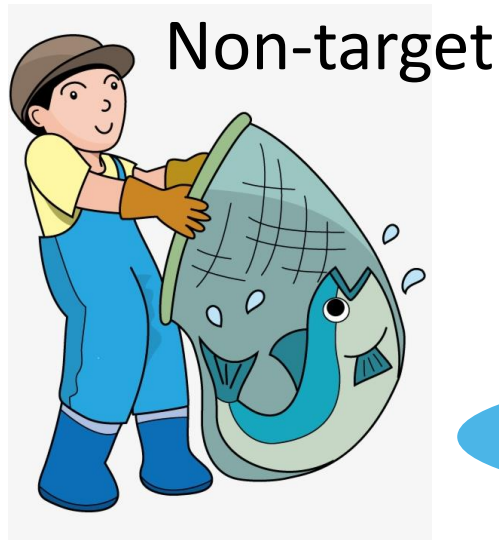
- Non-targeted analysis using high resolution mass spectrometry (HRMS) is a powerful approach for comprehensive chemical characterization in environmental samples
- **Interoperability** is the major challenge: vendor specific data formats and lab specific workflows hinder data sharing
- Overcoming this challenge positions the data for broad **Reusability** because it can be used to retrospectively identify novel contaminants
- No online repositories currently exist for these data, and ontologies are not fully developed.

This project seeks to demonstrate methods for sharing and analyzing HRMS data across instrumental and software platforms

Actions

- We've shared datasets of HRMS analyses. These are analyzed at the partner institution using established protocols and algorithms at Duke and UCD for compound annotation/identification from HRMS data, applying consistent performance metrics (e.g. confidence scale, mass accuracy, etc.)
- Currently, our data set struggles with **interoperability**.
 - Both groups have analyzed Hurricane Florence Water Pollution data, and observed disparate molecular features generated by open-source software.
 - Yurok Tribe data from Lumber Mill Pond was not acquired in MS method compatible with Duke University workflow.

Is there an ideal way process MS data?



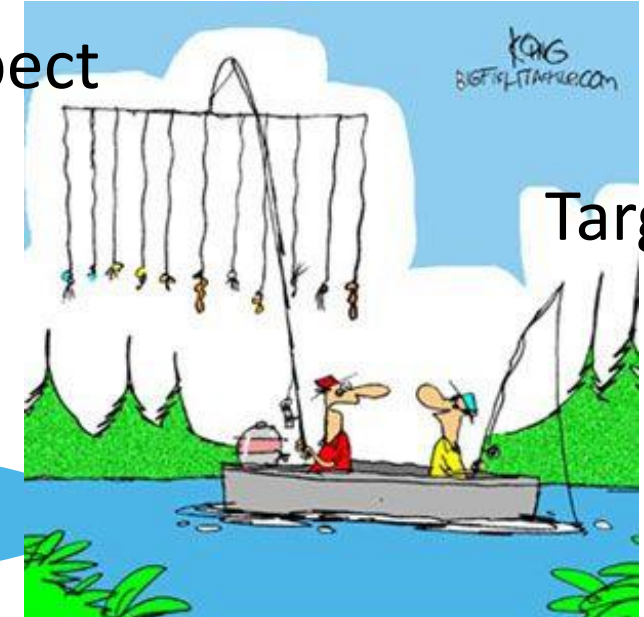
Non-target Screening

- Analytical choices aim so that the largest possible breadth of compounds are captured
 - Ex. Lab, Hardware, Software Choices to include compounds (rather than exclude them)

All Mass Spectrometry Data

Both share a need to define chemical space of analysis (much easier to define with target/suspect screening)

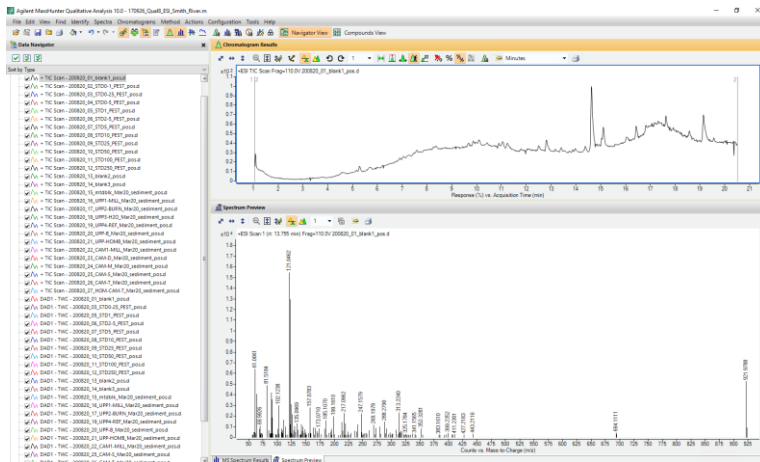
Suspect



Target/Suspect Screening

- Analytical choices aim for a group of compounds specific to study
 - Ex. Regulatory bodies use compound specific workflows to quantify compounds, etc.

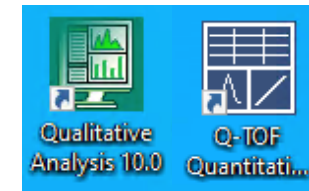
Workflow diagram



Raw data collected



Qualitative & Quantitative Analysis



Deconvoluted Peak picking (3d image)



Aligned peaks (consolidation)



MS-FLO

Mass Spectral Feature List Optimizer

MS-FLO is a tool to improve the quality of feature lists/peak tables after initial processing to expedite the process of data curation. It utilizes retention time alignments, accurate mass tolerances, Pearson's correlation analysis, and peak height similarity to identify ion-adducts, duplicate peak reports and isotopic features of the main monoisotopic metabolites. Removing such erroneous peaks reduces the overall number of metabolites in data reports and improves the quality of subsequent statistical investigations.

MS-FLO was developed at Fiehn Research Lab and has been recently published in Analytical Chemistry:

DeFolice BC, Mehta SS, Samra S, Čajka T, Wanciewicz B, Fahrman JF, Fiehn O. Mass Spectral Feature List Optimizer (MS-FLO): a tool to minimize false positive peak reports in untargeted LC-MS data processing. Analytical Chemistry, February, 2017. DOI: 10.1021/acs.analchem.6b04372

MS-FLO is open source and released under LGPL 3.0. If you encounter any bugs or have suggestions, please use the issue tracker.

Filtering steps



Data set ready to assign molecular formulas

What are FAIR ways that we can share MS data?

Sharing suspected compounds

1. Curate suspect list

- The compounds known to exist in a particular group of suspected contaminants
- Ex. NORMAN suspect list

2. Repository of MS on Database

- Mass spectrum of compound IDs are deposited into a database with appropriate metadata (analytical choices, instrumentation, etc)
- Ex. MONA

Sharing suspected data sets

3. Publish MS data from vendor specific or open-source format in journal, appendix, database, etc.

- MS for a specific dataset is published for public use.
- Ex. Appendix documents contain full alignment results

4. Share raw MS data of a suspect dataset with another lab for analysis

- One lab processes raw data collected by a different lab
- Ex. This project uses raw data collected from one lab to be processed by the other lab

What are FAIR ways that we can share MS data?

Advantage vs. Disadvantage

Sharing suspected compounds

1. Curate suspect list

- Advantage: Relatively easy to create / integrate into new workflow
- Disadvantage: Not thoroughly investigating chemical space

2. Repository of MS on Database

- Advantage: Ease of access for comparison to data generated by other labs
- Disadvantage: compounds are individually added into database; hard to get clean spectrum

Sharing suspected data sets

3. Publish MS data from vendor specific or open-source format in journal, appendix, database, etc.

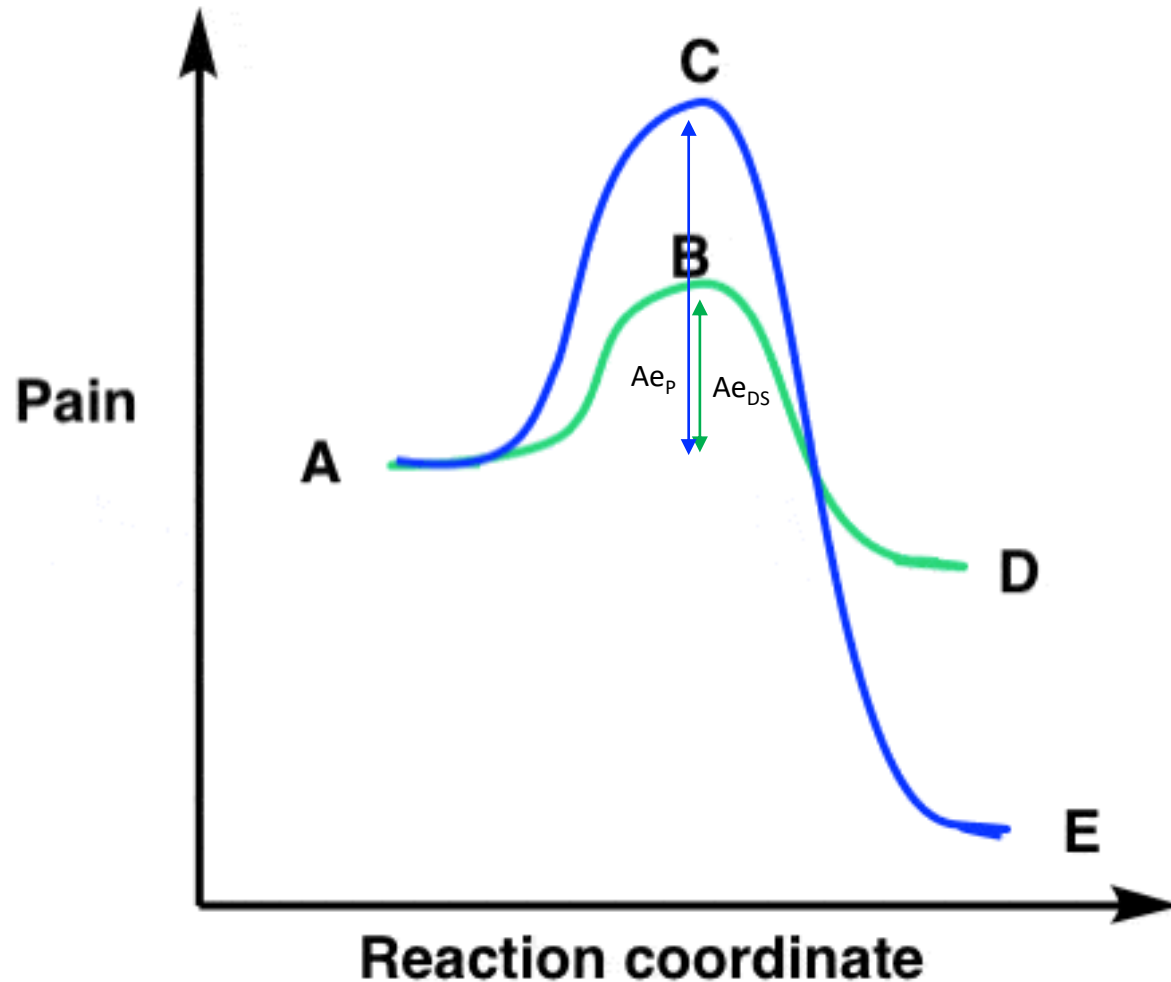
- Advantage: Relatively easy to publish output of software analysis
- Disadvantage: Limited in ability for other labs to process data (missing interoperability)

4. Share raw MS data of a suspect dataset with another lab for analysis

- Advantage: Data analysis not connected to a place or time, look at old data for new information
- Disadvantage: Making data interoperable is a major analytical consideration

In this example..
Activation energy = money

Reaction coordinate for can opener purchases



Legend (not to scale)

— Dollar Store Can Opener
— "Premium" Can opener

A: No can opener

B: Buying a crappy can opener

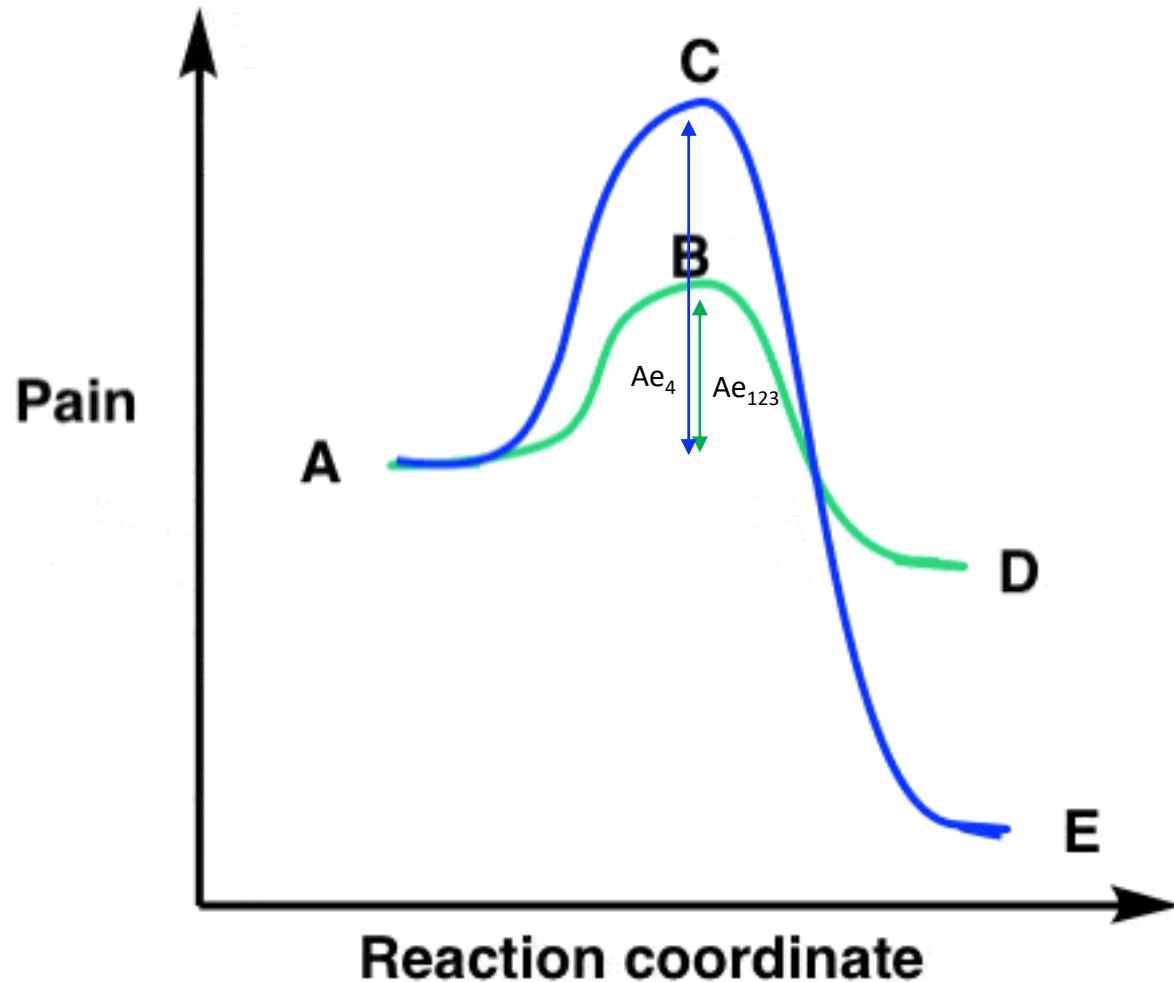
C: Buying a premium can opener
(painful!)

D: Owning a crappy can opener



E: Owning a premium can opener

Reaction coordinate for MS Data Sharing

In this example..
Activation energy = interoperability



Legend (not to scale)

	Option 1, 2, 3
	Option 4

A: Not sharing MS Data

B: Sharing MS Data via option 1, 2, 3

C: Sharing MS Data via option 4
(painful!)

D: MS Data that another lab processed

E: MS Data that your lab processed

Because option 1, 2, 3 are easier to achieve, they are often the output

Example: Yurok Tribe Lumber Mill Site Data

Sharing suspected compounds

1. Curate suspect list

- Use curated list to identify compounds thought to exist at lumber mill sites
- Create suspect list of biologically active compounds from bioassay results

2. Repository of MS on Database

- Contribute compounds successfully identified by workflow into MS database

Sharing suspected data sets

3. Publish MS data from vendor specific or open-source format in journal, appendix, database, etc.

- Easy to contribute but little chance for interoperability
- Reflects analytical decisions from Young lab sample processing

4. Share raw MS data of a suspect dataset with another lab for analysis

- Labs with different expertise can analyze data using different workflow
- Difficult to make interoperable, analytical choices hinder perfect compatibility

Next Steps...

Describe Chemical Space

1. Describe known chemical space of data sets – physical & chemical properties of target and library matched compounds
 - MW and m/z
 - Log_{kow}
 - pK_a
 - Ionization
 - uses
2. Describe chemotype coverage of methodology and analysis for both data sets
3. Development of target training sets of compounds to identify via target/suspect screening (25 - 50 compounds)
4. Development of non-target training sets of compounds to identify via molecular formula generation (25 - 50 compounds)

Make Data FAIR: Interoperability

1. Address interoperability of the Yurok Tribe Lumber Mill Site data.
 - If unable to proceed with data independent analysis (DIA) in MS method, samples must be re-analyzed.
2. Address interoperability of the UCD-Duke workflows.
 - Filtered and processed data showed disparate results for feature annotation.
 - Use online Venn diagram tools to describe overlapping features between each workflow