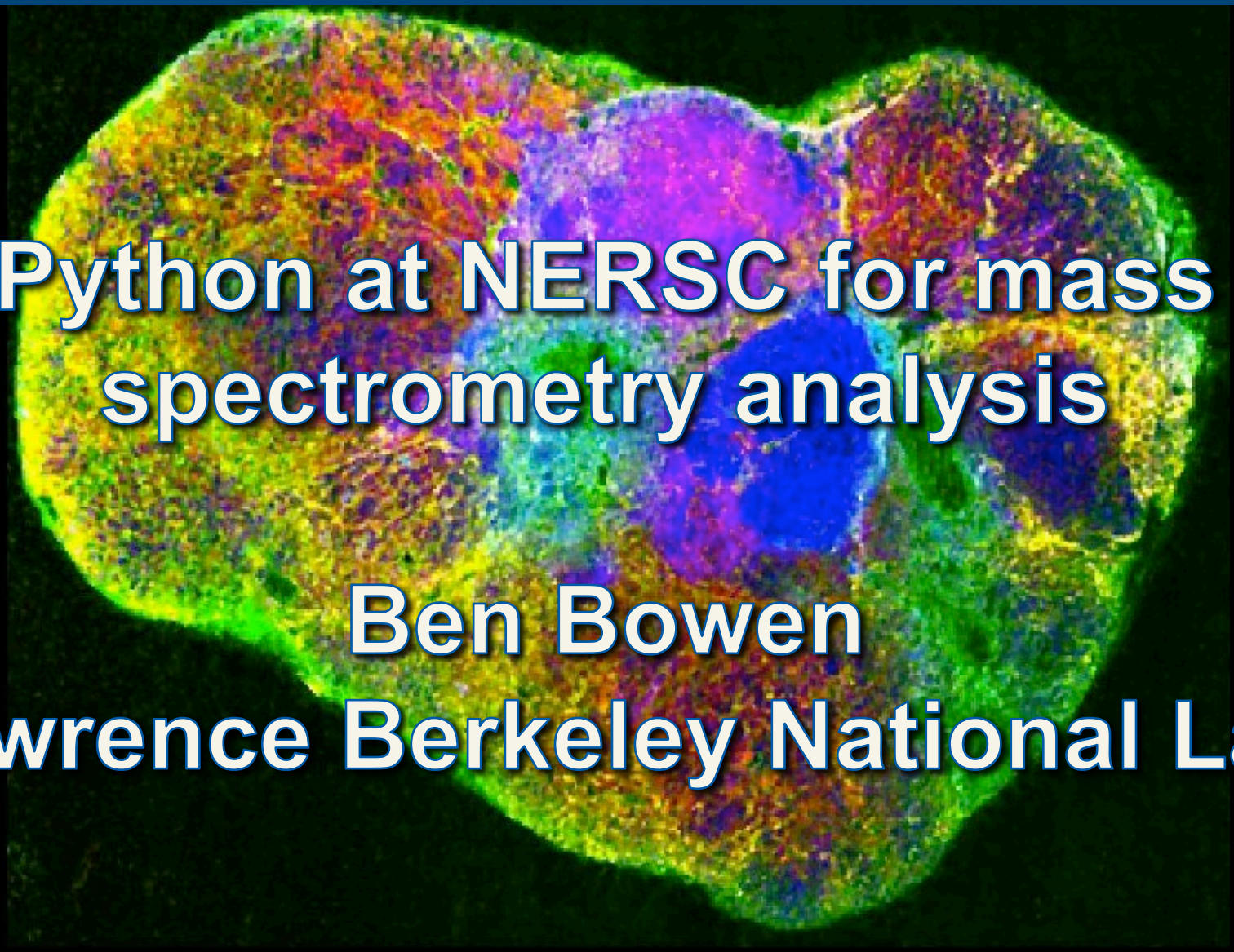
BERKELEY LAB

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY



Python at NERSC for mass spectrometry analysis

Ben Bowen
Lawrence Berkeley National Lab

Who am I? Where I work? Why I use NERSC?

BERKELEY LAB

https://github.com/biorack
https://github.com/benbowen
http://openmsi.nersc.gov

BPBOWEN@LBL.GOV
Twitter: @bpbwc

Impact

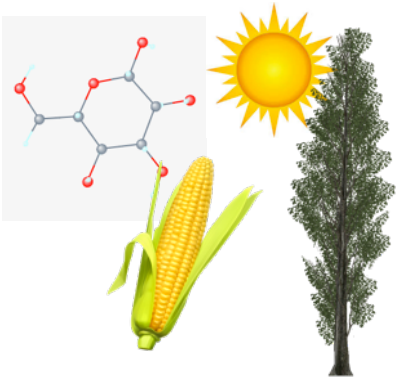
Community of users uploading and analyzing their own data and sharing their learning

(JGI)

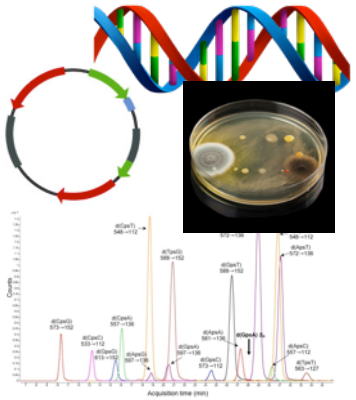
Illuminating the dark regions of biochemical space

BERKELEY LAB

Renewable Feedstocks



Pathway Engineering



Industrial Scale-up

Benefit

- Agriculture
- Fuels
- Sensors
- Food
- Medicine
- Electronics
- Materials
- Personal Care
- Bioremediation
- Mining

How would this story change if we knew the sequence and function of all biochemical pathways and had reliable analytical tools to identify their products?

Pathway synthesis, coupled with ultrasensitive high-throughput metabolomics for identification of unknown functions

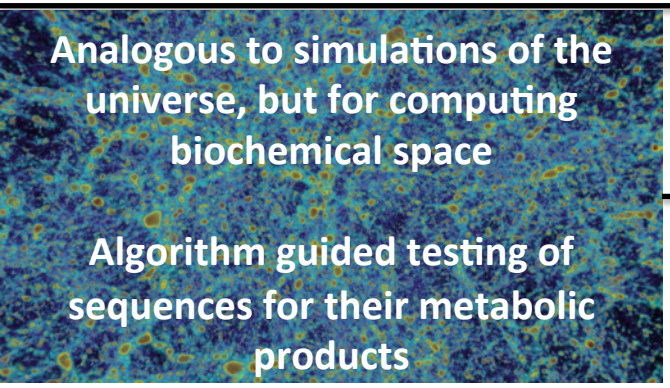
Advanced computing for metabolite identification, evidence based gene annotations, and efficient navigation of biochemical space

A fundamental understanding of biochemical space where little or no knowledge currently exists

Reliable design of biochemical systems with never before considered functions (BioFoundry)

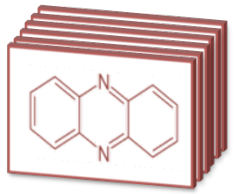
Advanced computing for metabolite identification

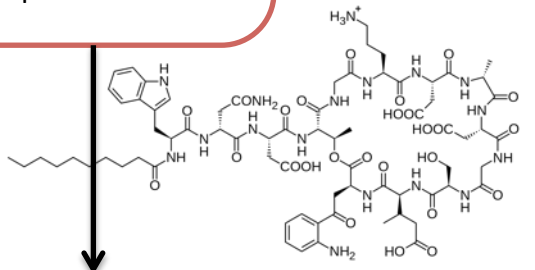
BERKELEY LAB



A new generation of advanced metabolite identification methods at the scales needed

Machine learning computational methods combined with state-of-the-art physical chemistry predictions

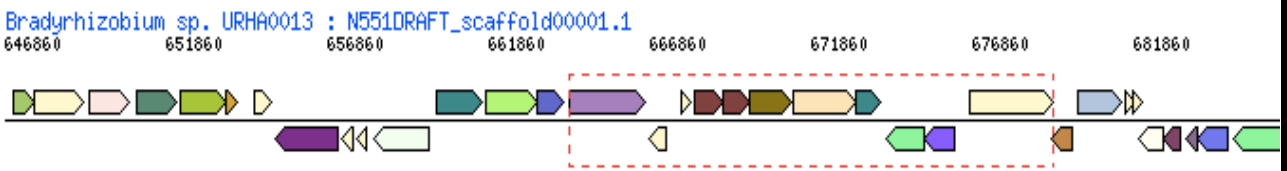


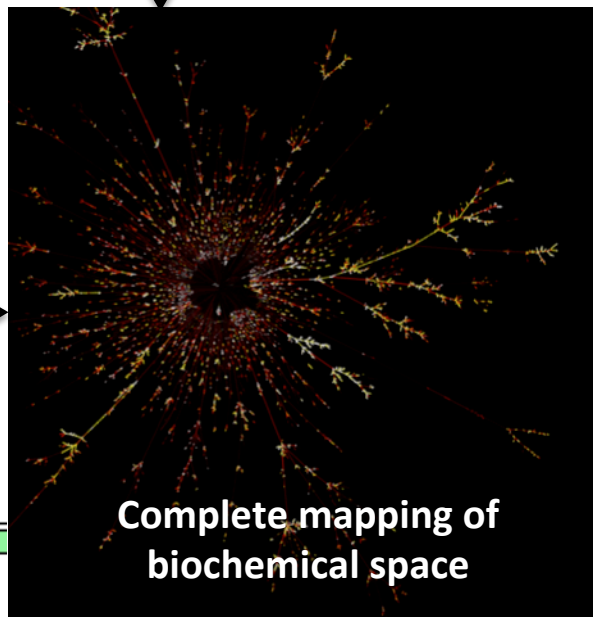


Subsequent steps are informed by the previous work and guided in a way to maximize discovery

Evidence based annotation and prioritization of high-novelty biosynthetic clusters

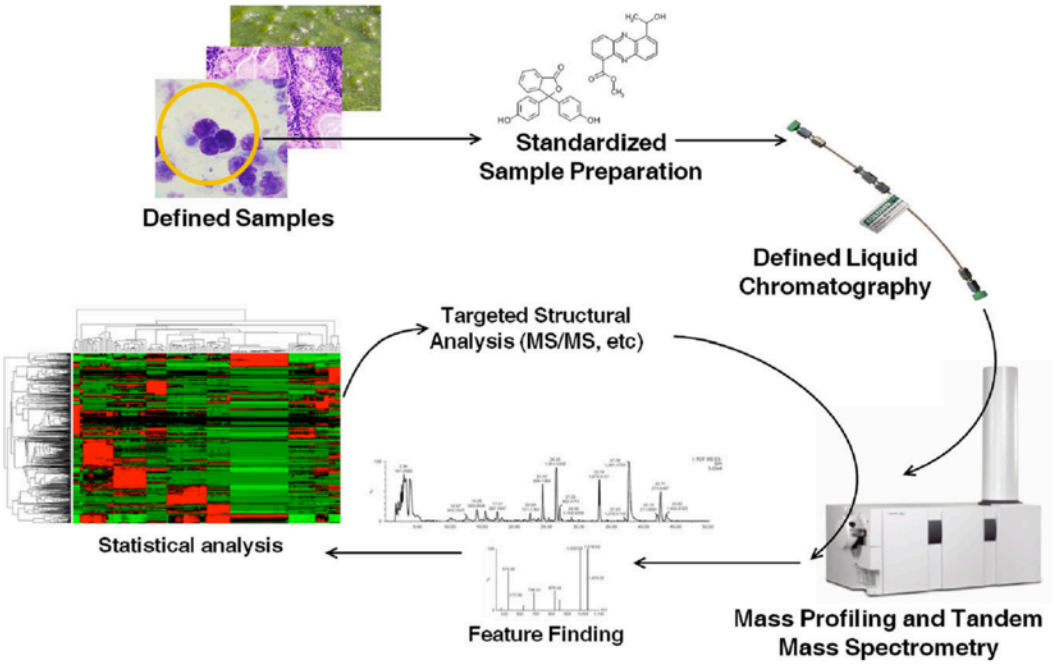
Scalable data access and management

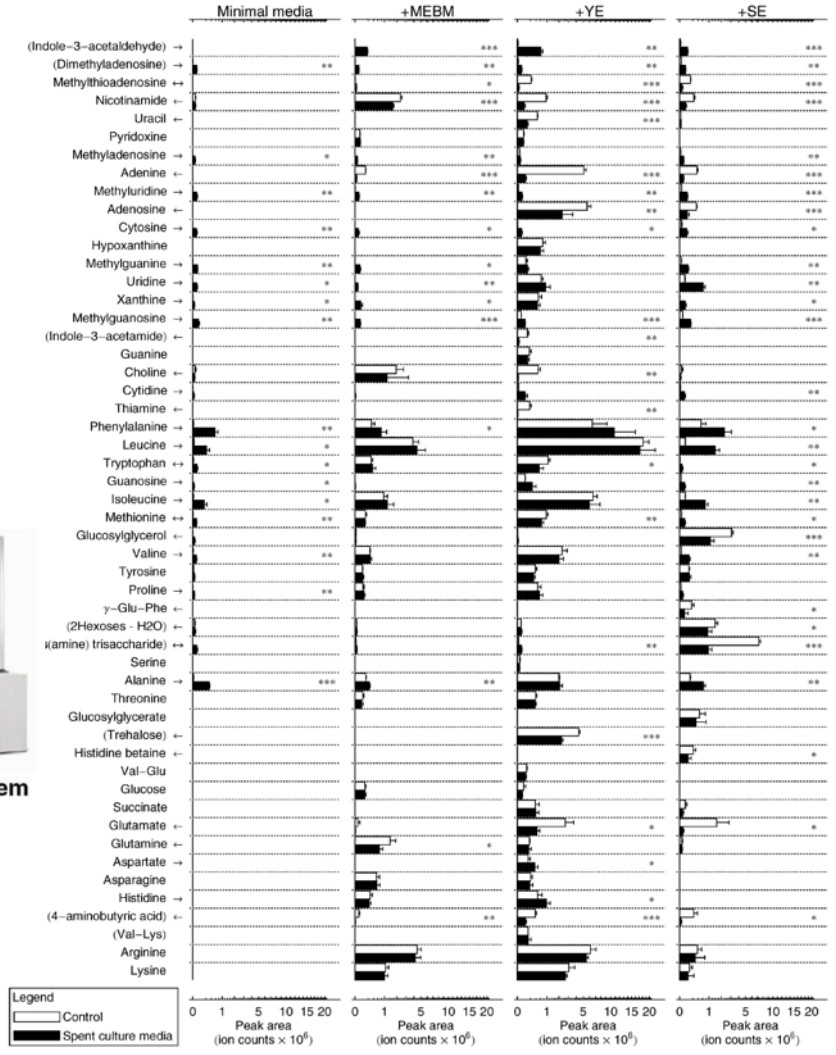




MS is an established metabolomics tool

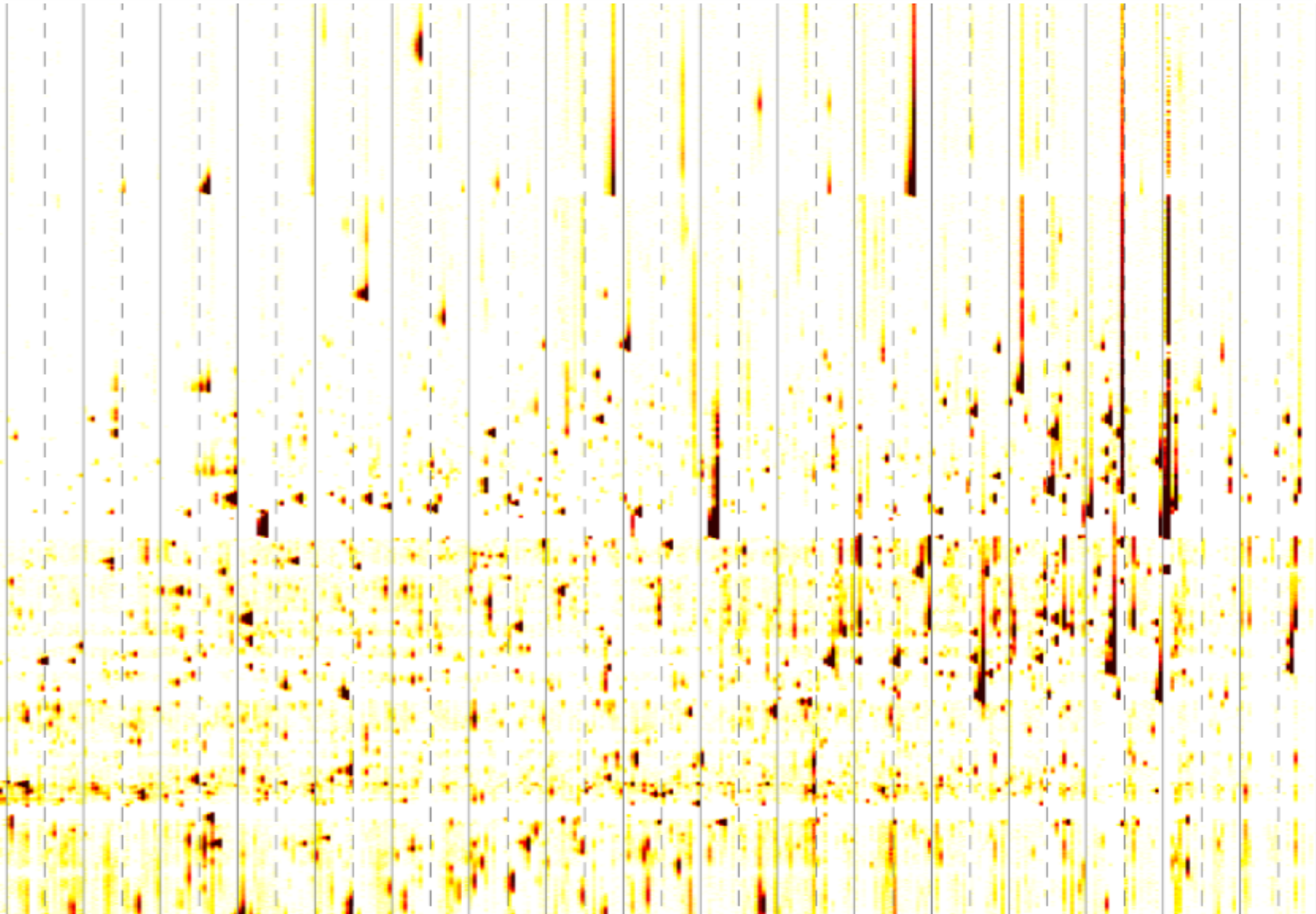
- LC-MS provides very high chemical resolution





Baran, R.; Bowen, B. P.; Northen, T. R. Untargeted metabolic footprinting reveals a surprising breadth of metabolite uptake and release by Synechococcus sp. PCC 7002. *Mol Biosyst* **2011**, *7*, 3200–3206.

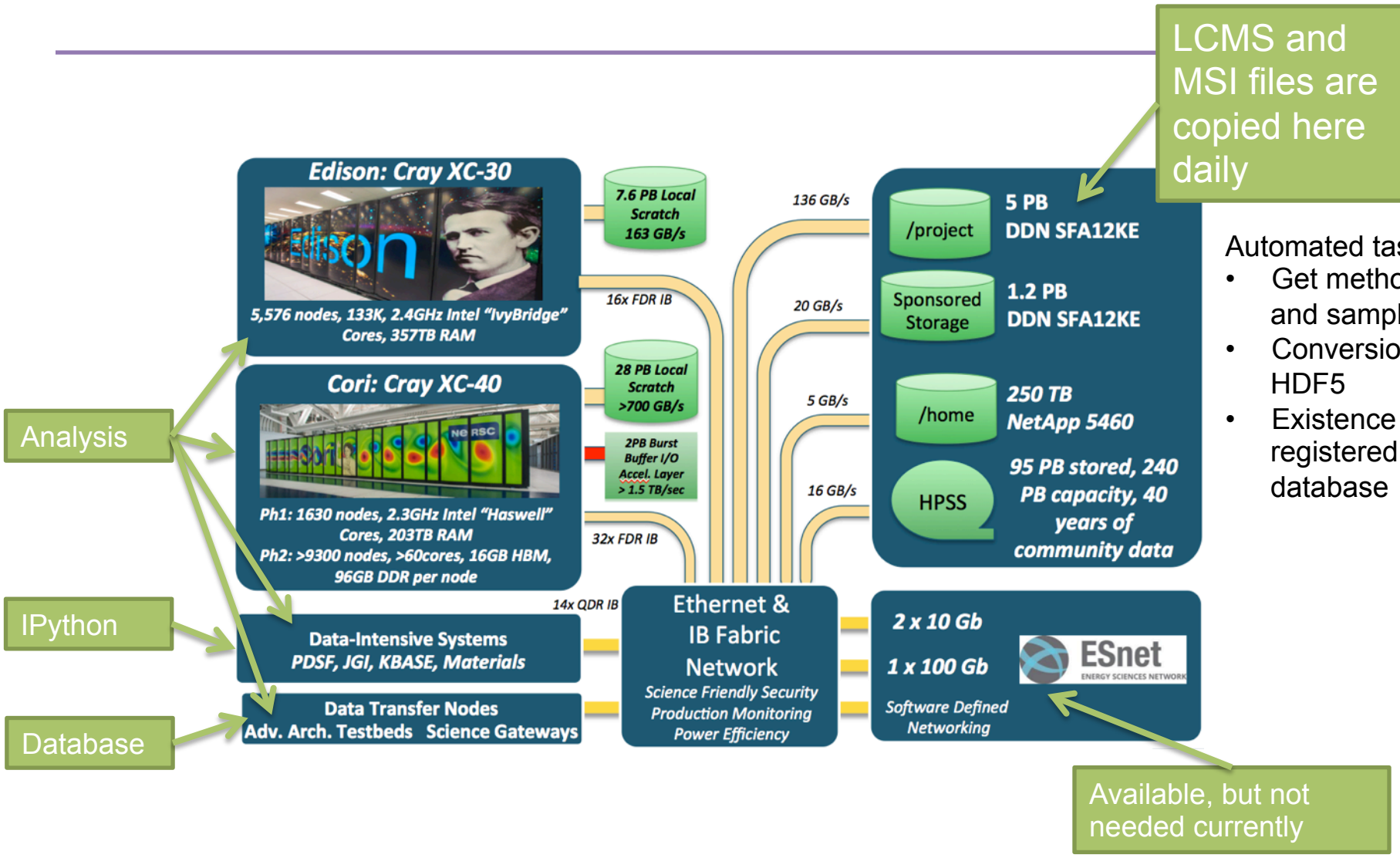
LCMS data is also a multimodal hyperspectral image



NERSC Computational Environment & MS integration

BERKELEY LAB

Quick and Painless: Manage, Analyze, Visualize LCMS and MSI data



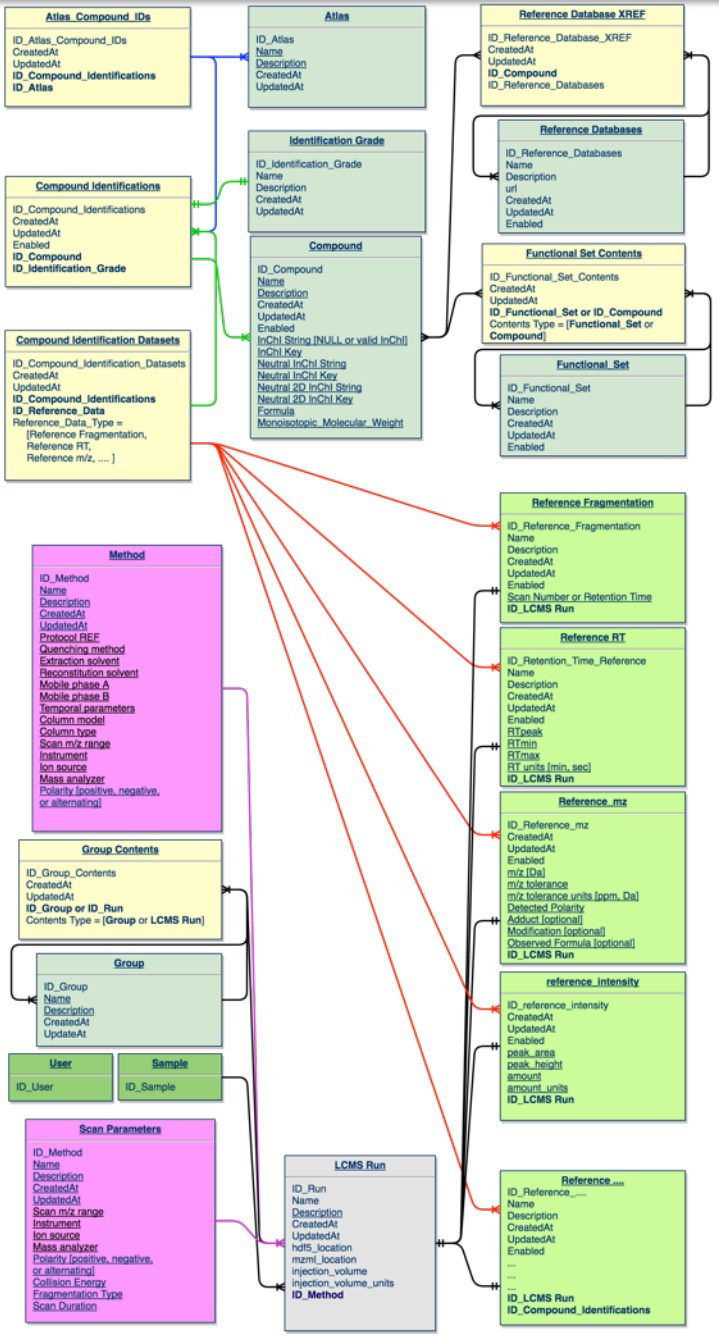
Automated tasks:

- Get method and sample
- Conversion to HDF5
- Existence is registered in database

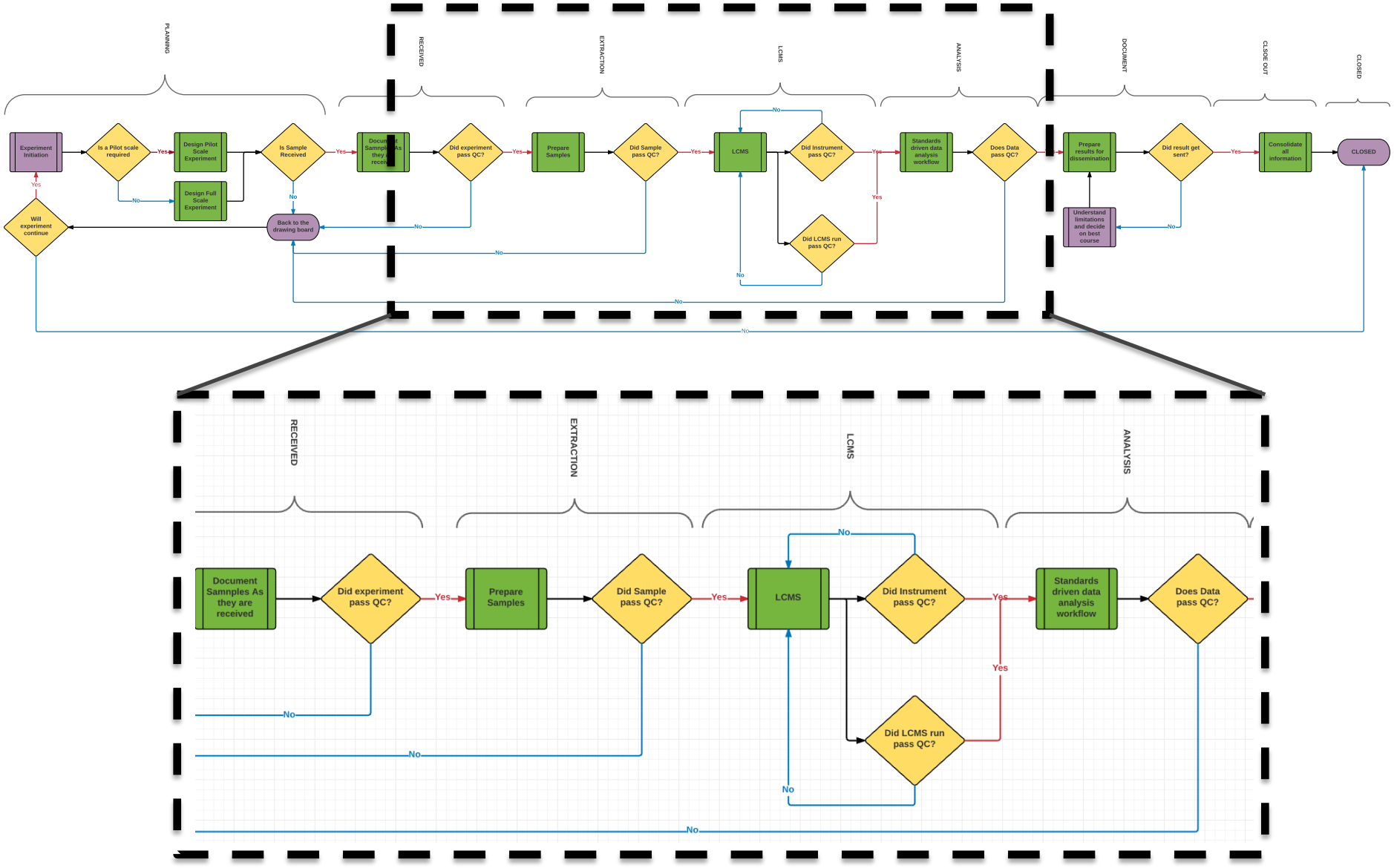
Metabolomics, HPC, and databases integrated

- A database to store assertions
- Maintain information to accelerate future experiments
- Metabolite discovery tools

1. Bowen, B. P., & Northen, T. R. (2010). Dealing with the unknown: metabolomics and metabolite atlases. *Journal of the American Society for Mass Spectrometry*, *21*(9), 1471–1476.
2. Fischer, C. R., Ruebel, O., & Bowen, B. P. (2016). An accessible, scalable ecosystem for enabling and sharing diverse mass spectrometry imaging analyses. *Archives of Biochemistry and Biophysics*, *589*, 18–26.
3. Wang, Y., Kora, G., Bowen, B. P., & Pan, C. (2014). MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Analytical Chemistry*, *86*(19), 9496–9503.
4. Yao, Y., Bowen, B. P., Baron, D., & Poznanski, D. (2015a). SciDB for High-Performance Array-Structured Science Data at NERSC. *Computing in Science & Engineering*, *17*(3), 44–52.
5. Yao, Y., Sun, T., Wang, T., Ruebel, O., Northen, T., & Bowen, B. P. (2015b). Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. *Metabolites*, *5*(3), 431–442.



Metricized checkpoints for improving efficiency and quality



MS-Monitor checks every LCMS Run

LCMSRUN(s) in MetAtlas Database

Get MS-Monitor Notebook

Module 1. Import Packages

Module 2. Select Your Experiment

Module 3. Retrieve the files for that experiment

Module 4. Specify the strings you used in your filenaming

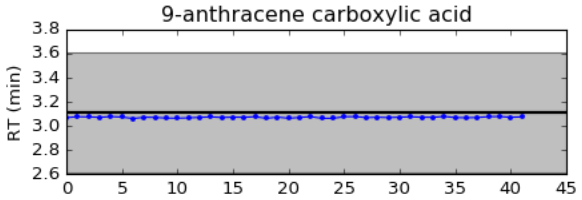
Module 5. Retrieve reference data for ISTDs and QCs

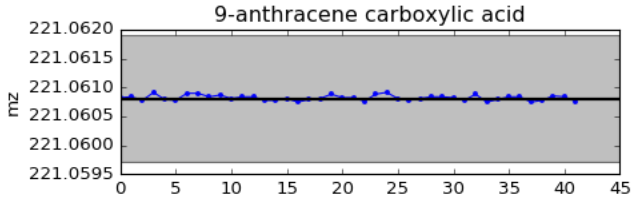
Module 6. Check the parameters

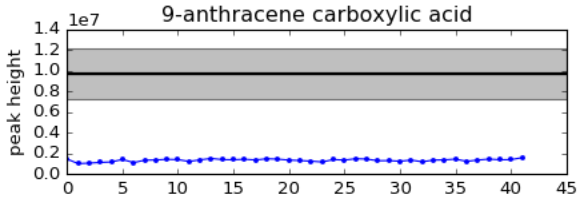
Module 7. Collect QC and ISTD data for each file, make table, and make figures

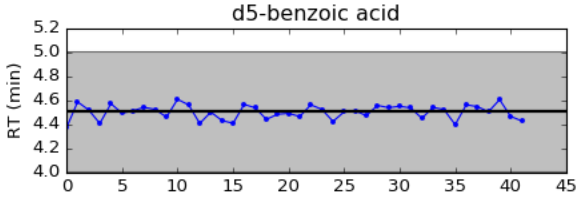
Module 8. Specify each lcmsrun's "pass_qc" attribute as True/False

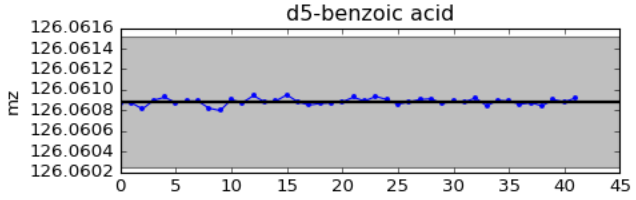
User needs to know which to re-run

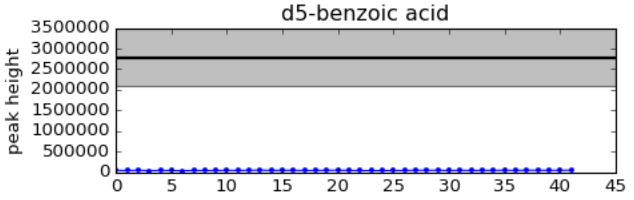


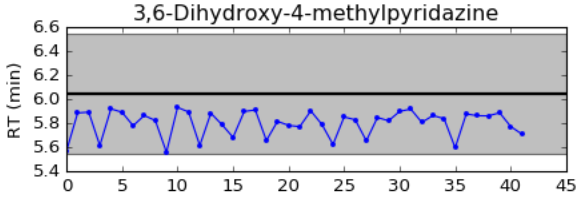


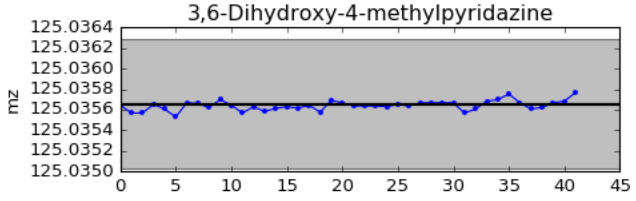


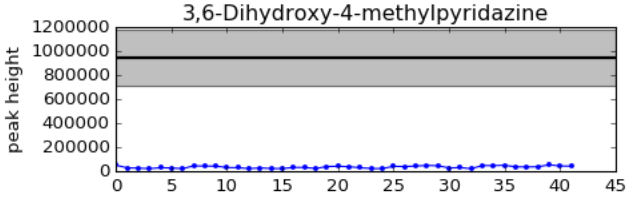


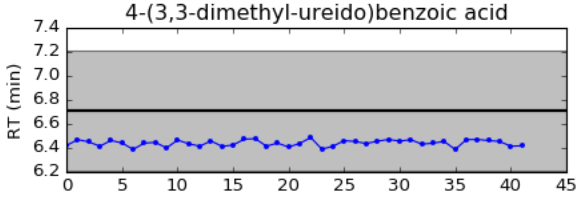


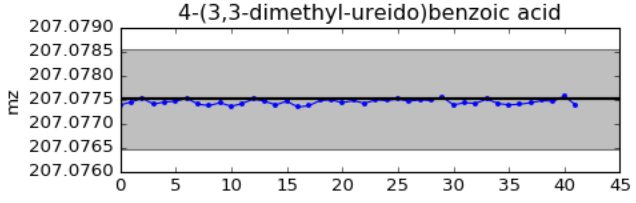


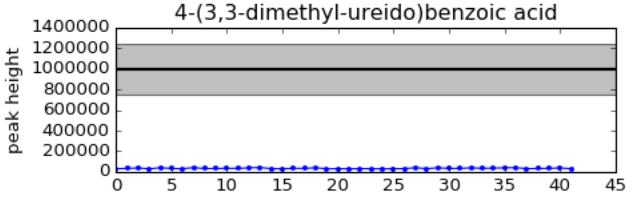


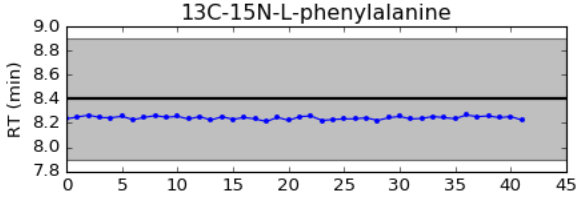


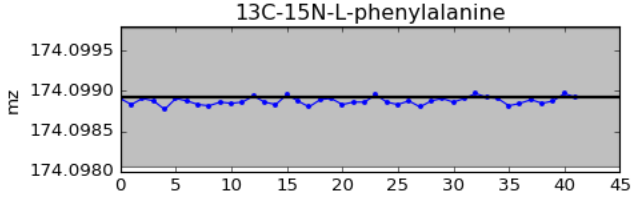


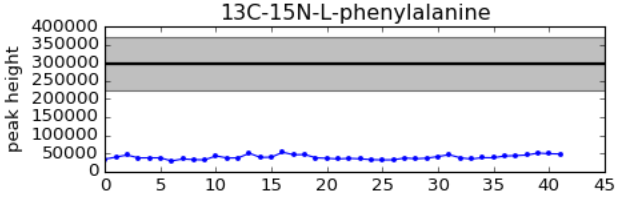


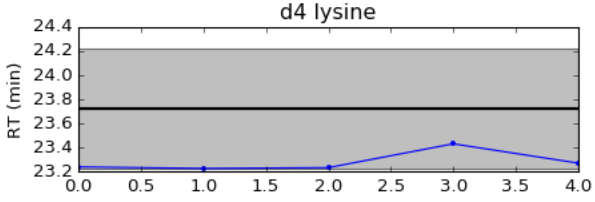


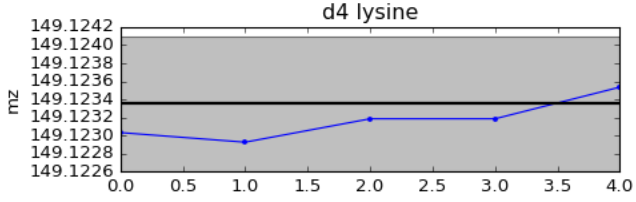


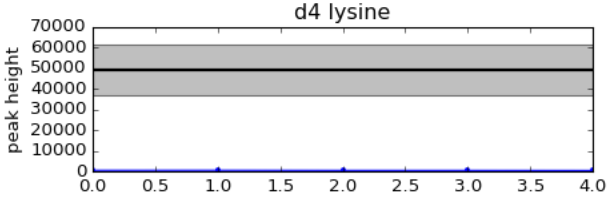






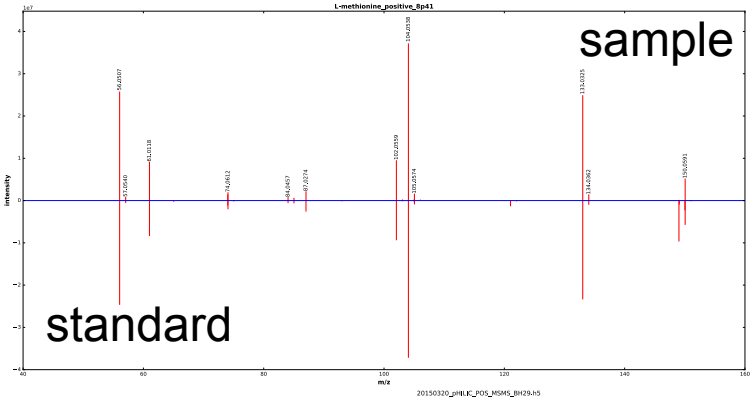


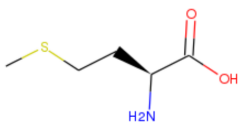


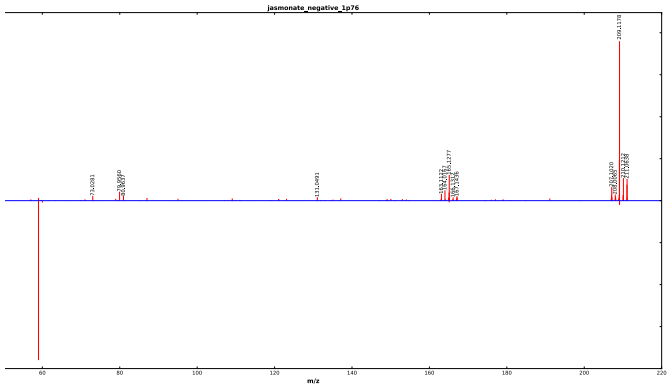


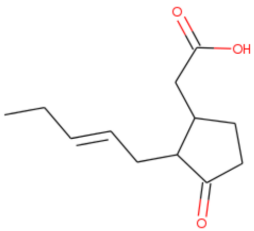
Standard based identification using MS/MS

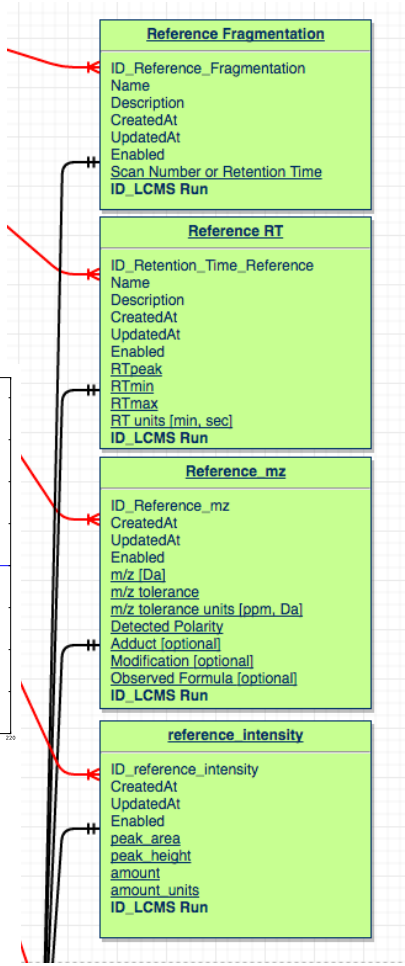
Layered identification to facilitate untargeted detection of intense ions











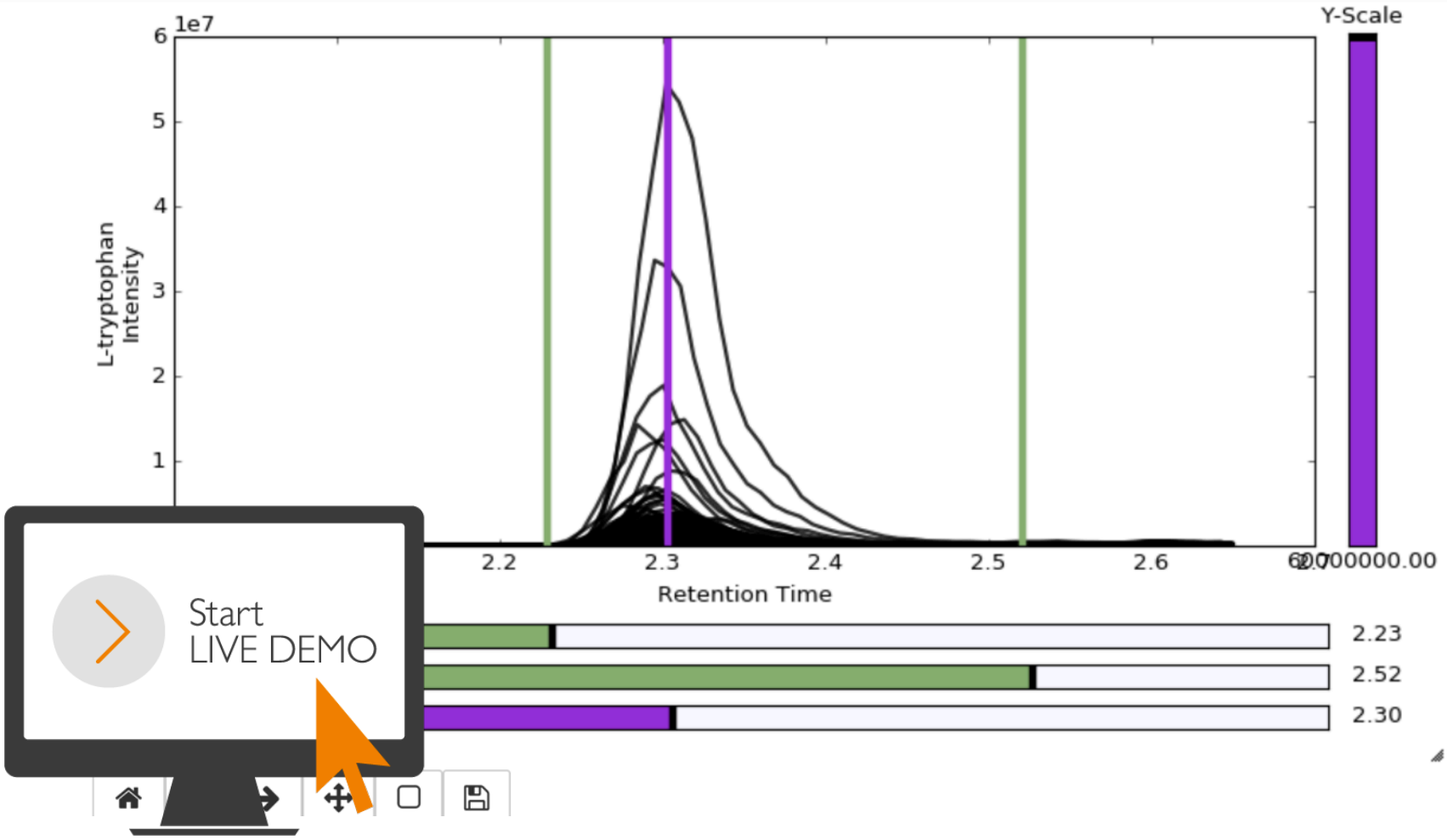
Retention Time and m/z reference

BERKELEY LAB

jupyter Workflow Notebook Last Checkpoint: Last Friday at 12:15 PM (unsaved changes)

File Edit View Insert Cell Kernel Help

Markdown CellToolbar



Dealing with the unknown

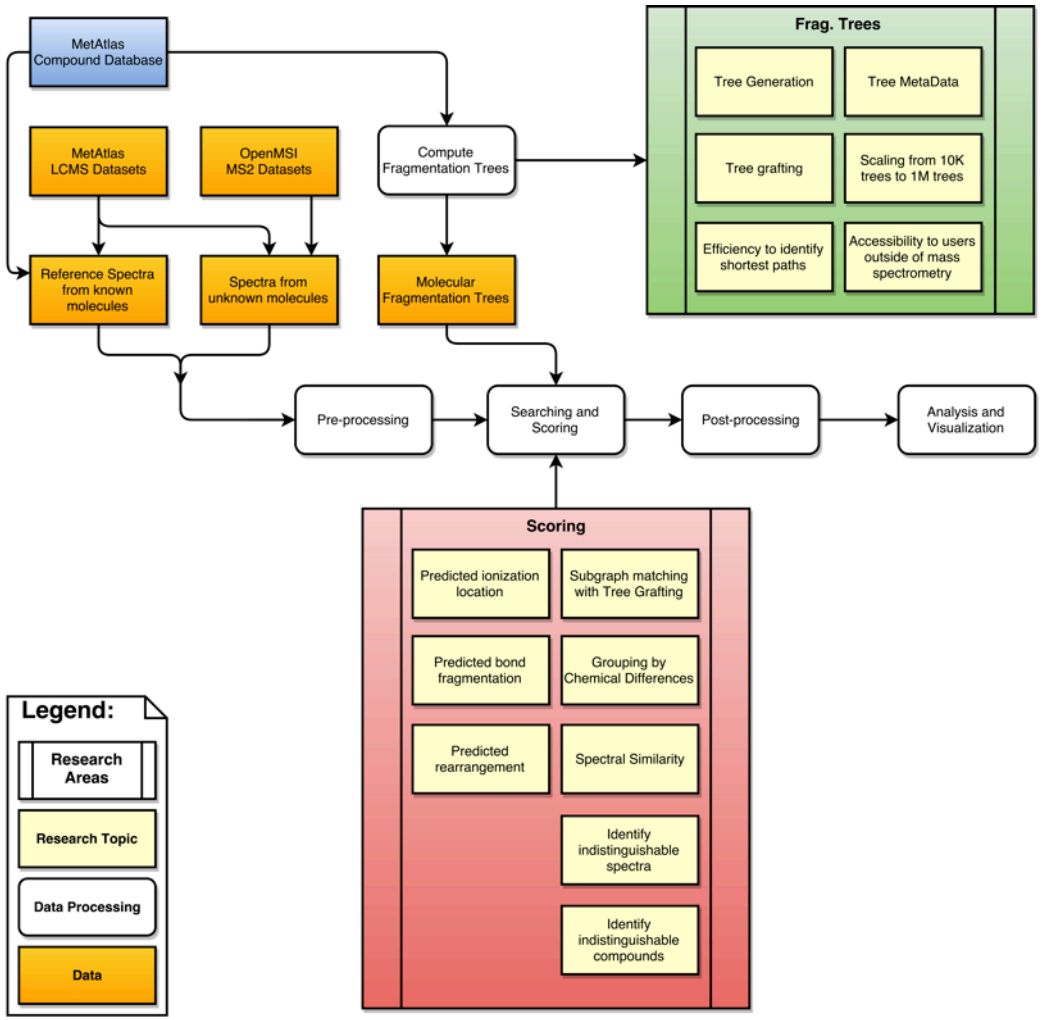
BERKELEY LAB

Compounds you observe: authentic standard is not available

1. First choice is always third party spectral libraries:
- Not downloadable
- Low quality
- Compound not available

2. First principles calculation
- Working for EI
- ESI still has a way to go

3. Hybrid methods
- Dependent on large graphs
- Optimization of scoring algorithms needed



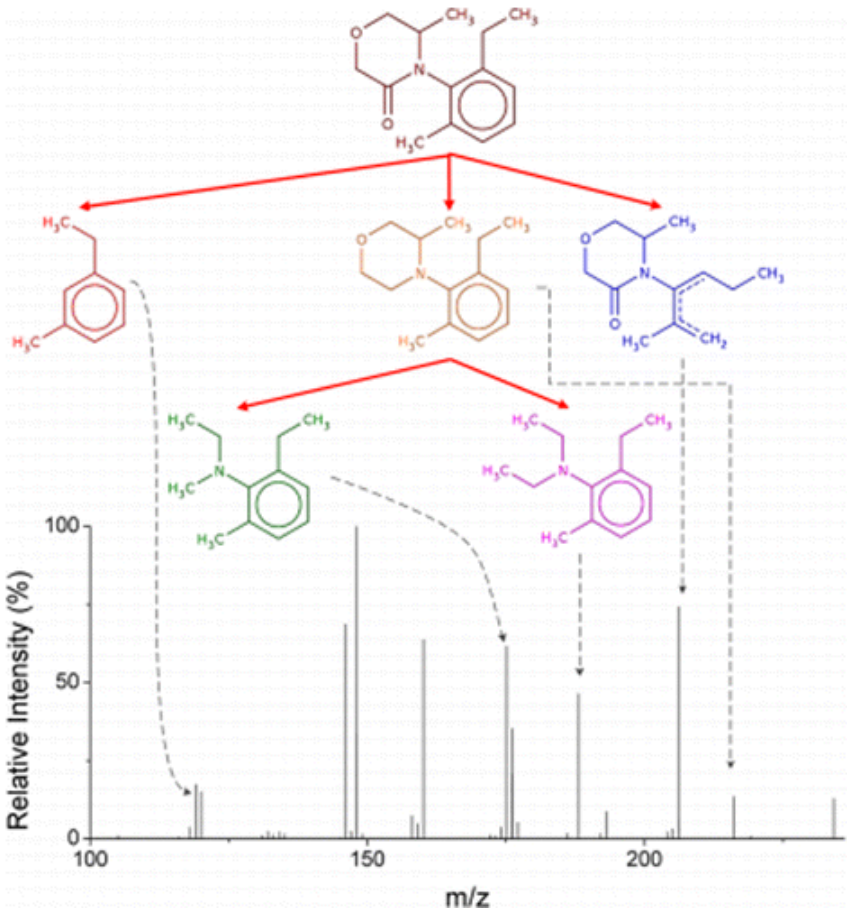
Tree based methods

Compounds you observe: authentic standard is not available

- Complete enumeration
- Scoring is higher when branch of tree is observed consistently

Much higher FDR than spectral libraries

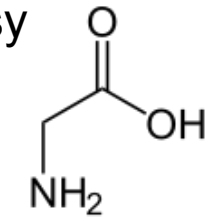
Enables exploration of hypothetical molecular structures which are not in databases



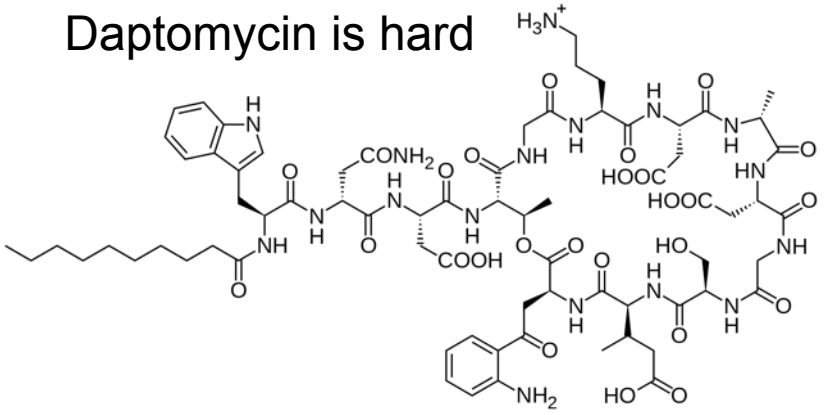
Access HPC from notebook interface

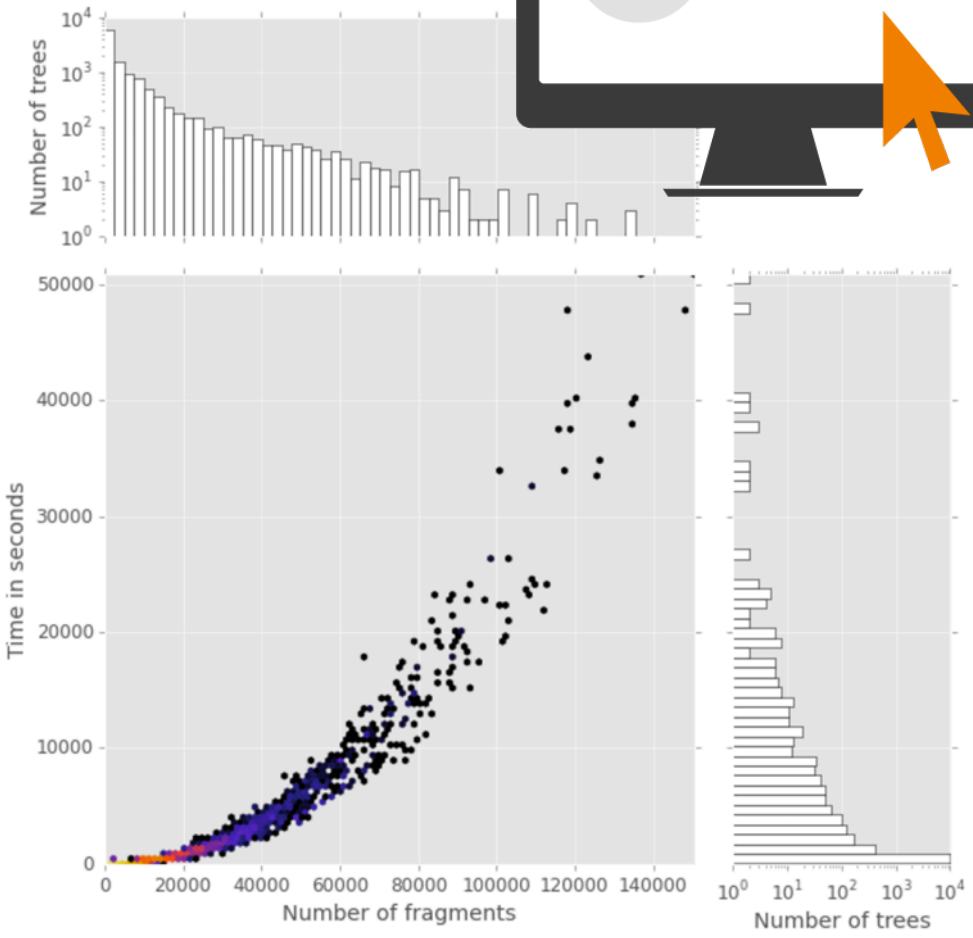
Without supercomputing this is not possible

Glycine is easy



Daptomycin is hard



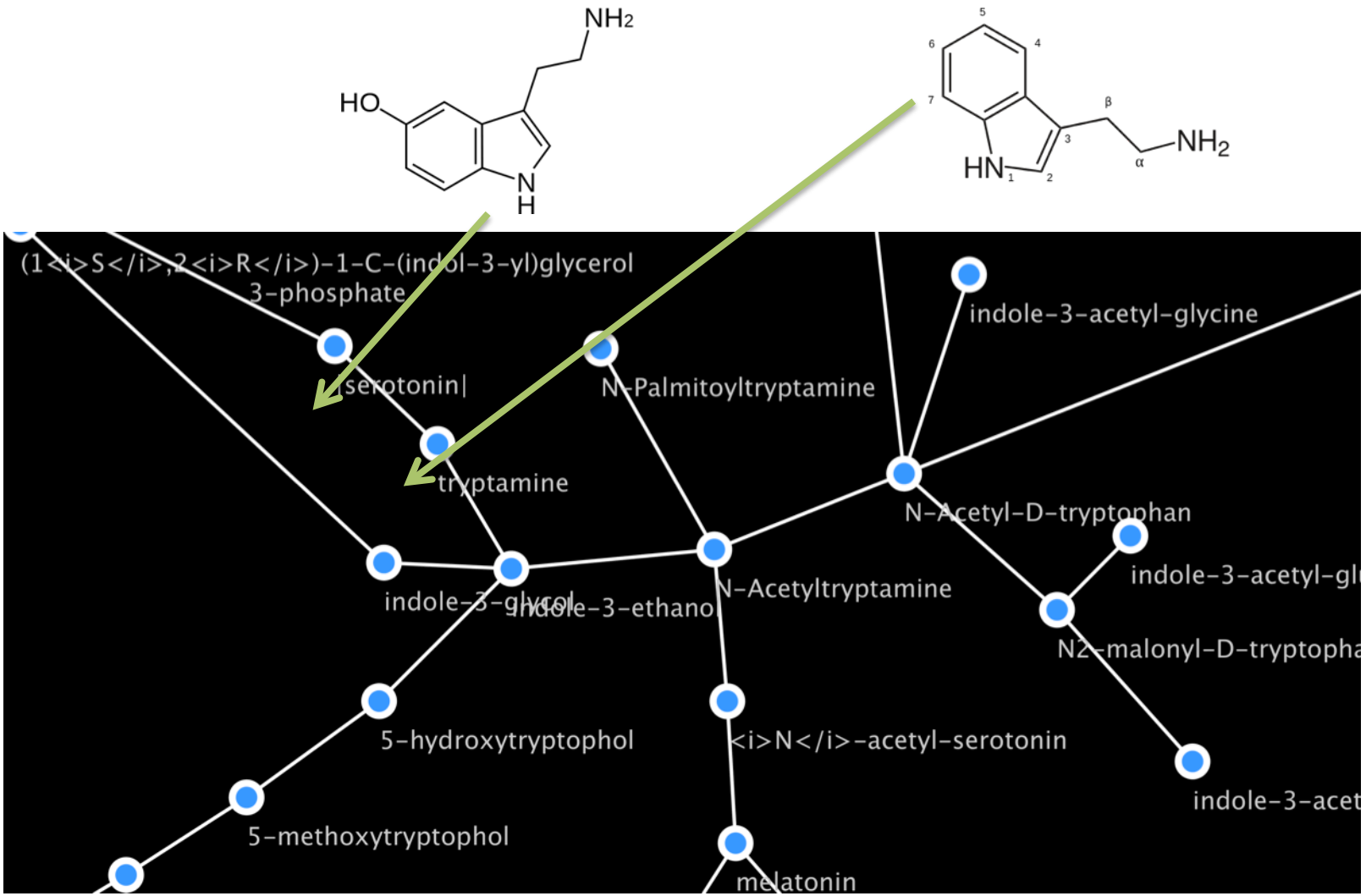


Data integration

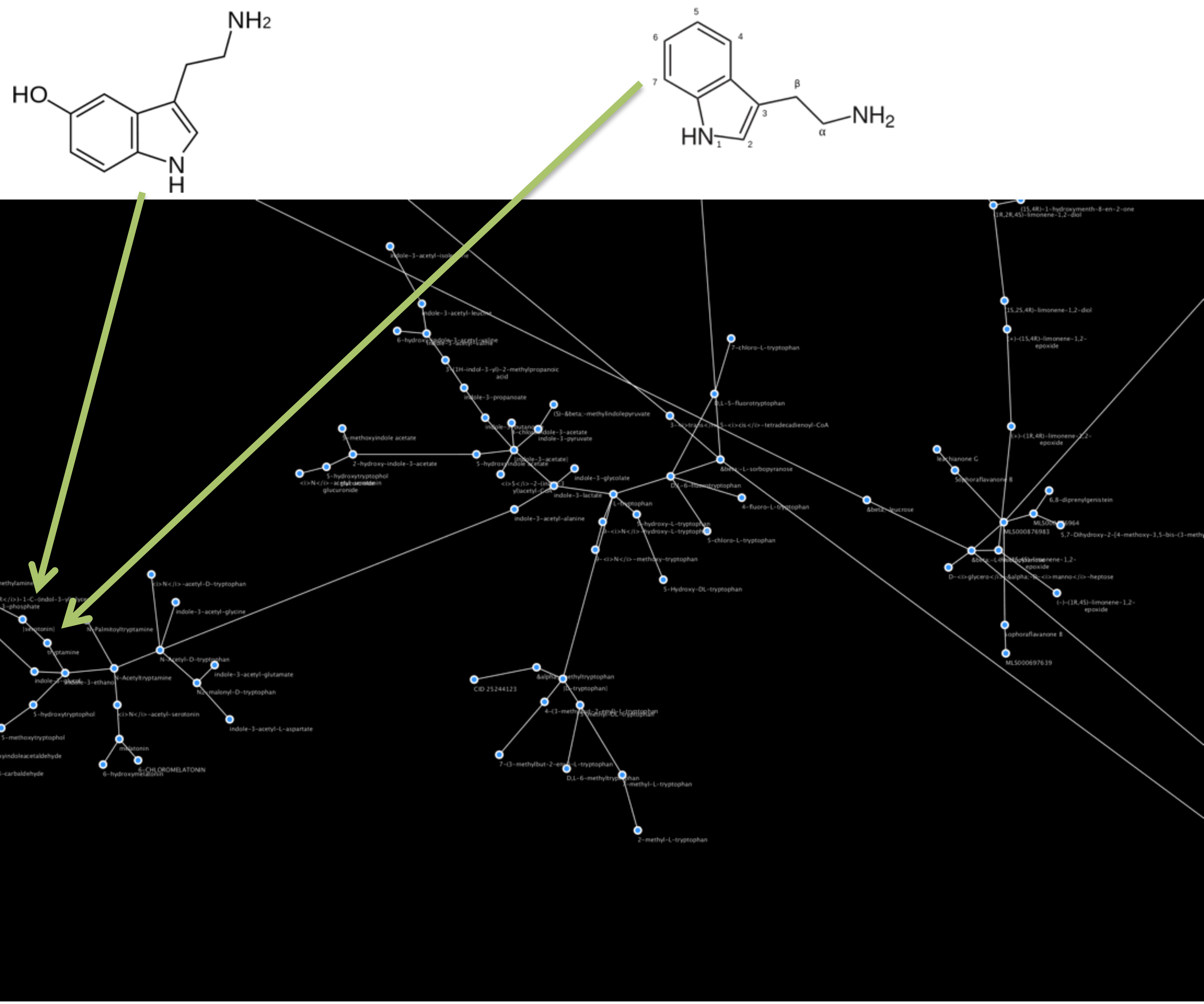
Visualize results of simultaneous measurement of many metabolites

- Consider a study where
  - Consumption of medium #1 induces sleep
  - Consumption of control media has no effect
- Identify the patterns of metabolite utilization
  - Find co-varying similar metabolites
  - Find classes of compounds

Chemical networks

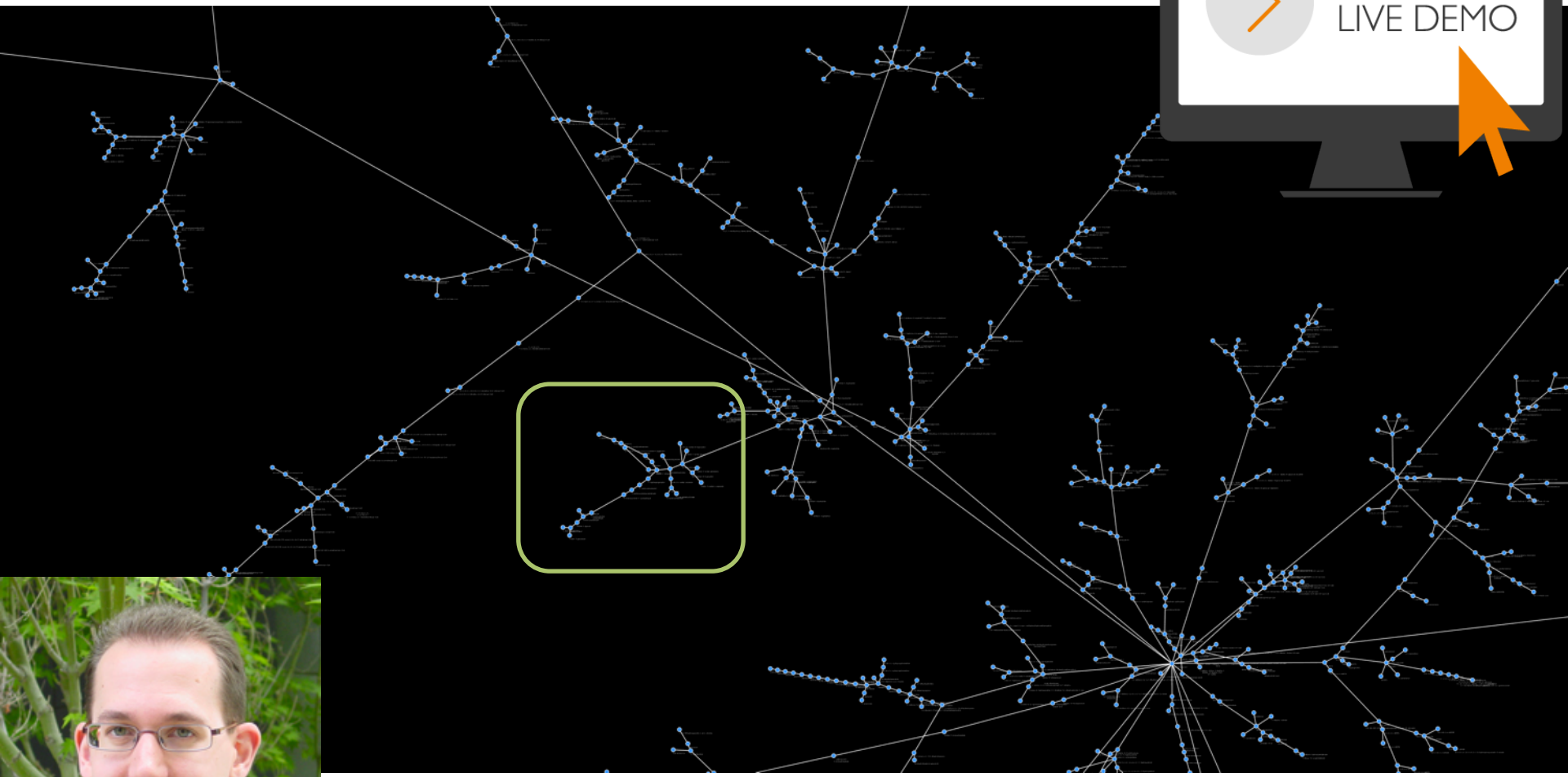


Chemical networks



Chemical networks

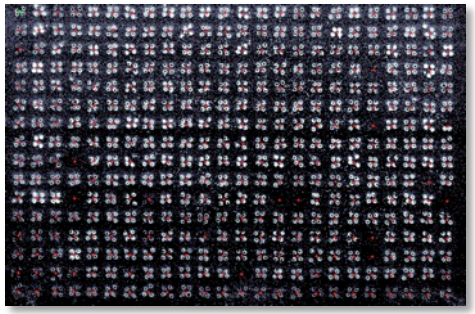
BERKELEY LAB



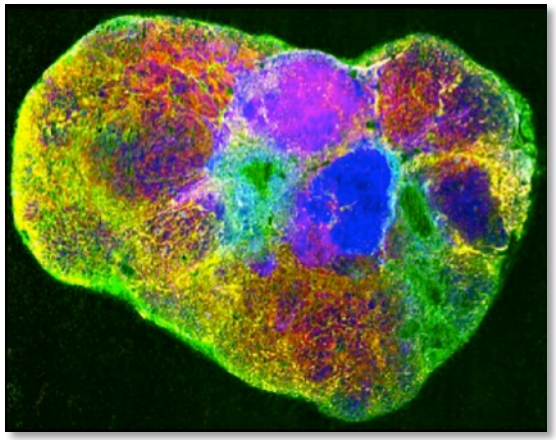
Widely used to measure chemical images

- Spatial gradients important in many applications

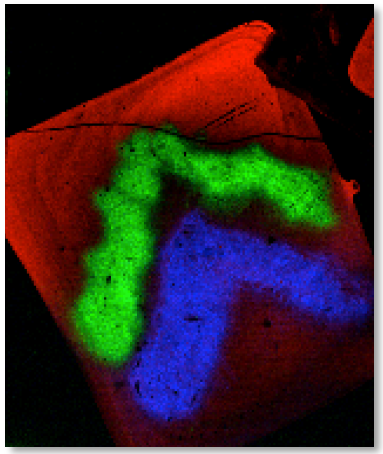
High-throughput screening



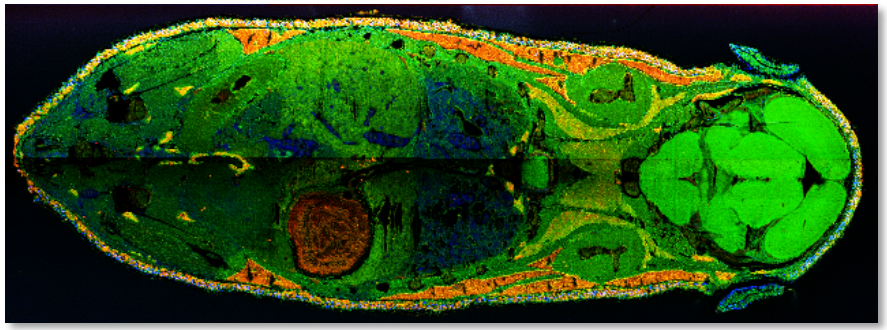
Pathophysiology



Microbial Interactions

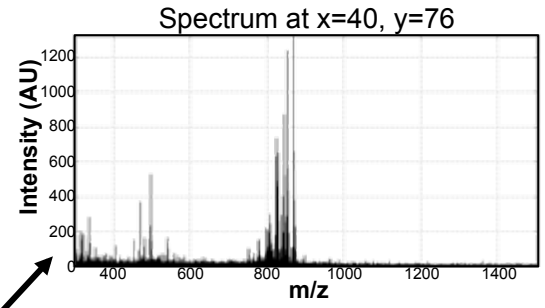


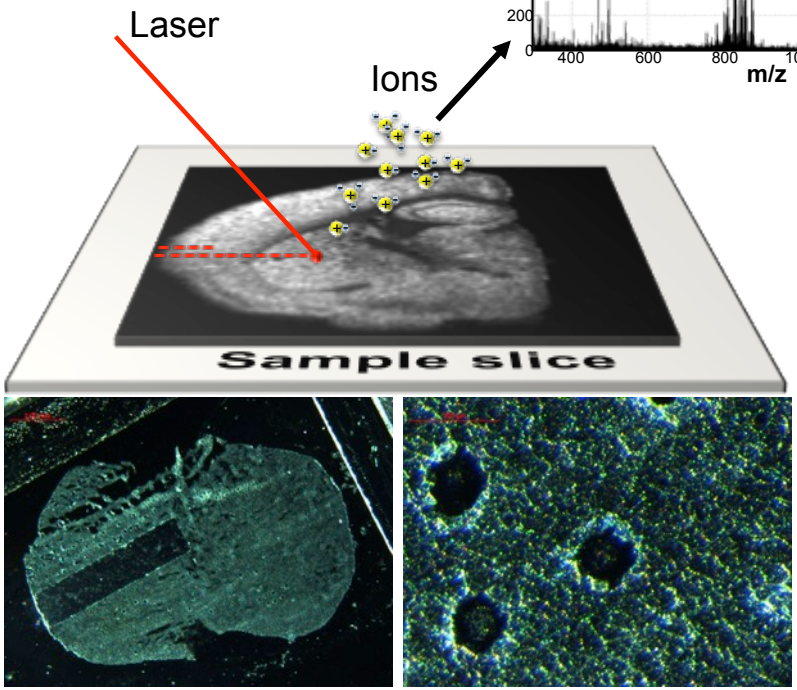
Drug metabolism



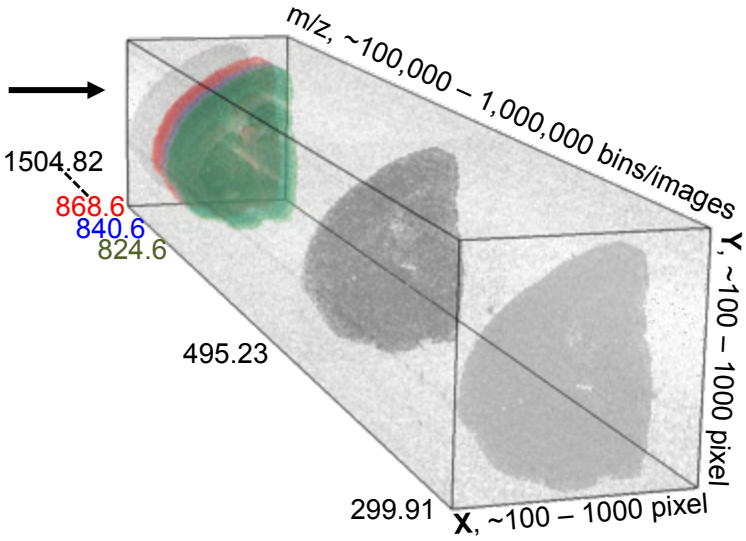
Mass spectrometry imaging

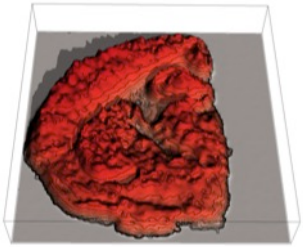
Each pixel has a complex spectrum





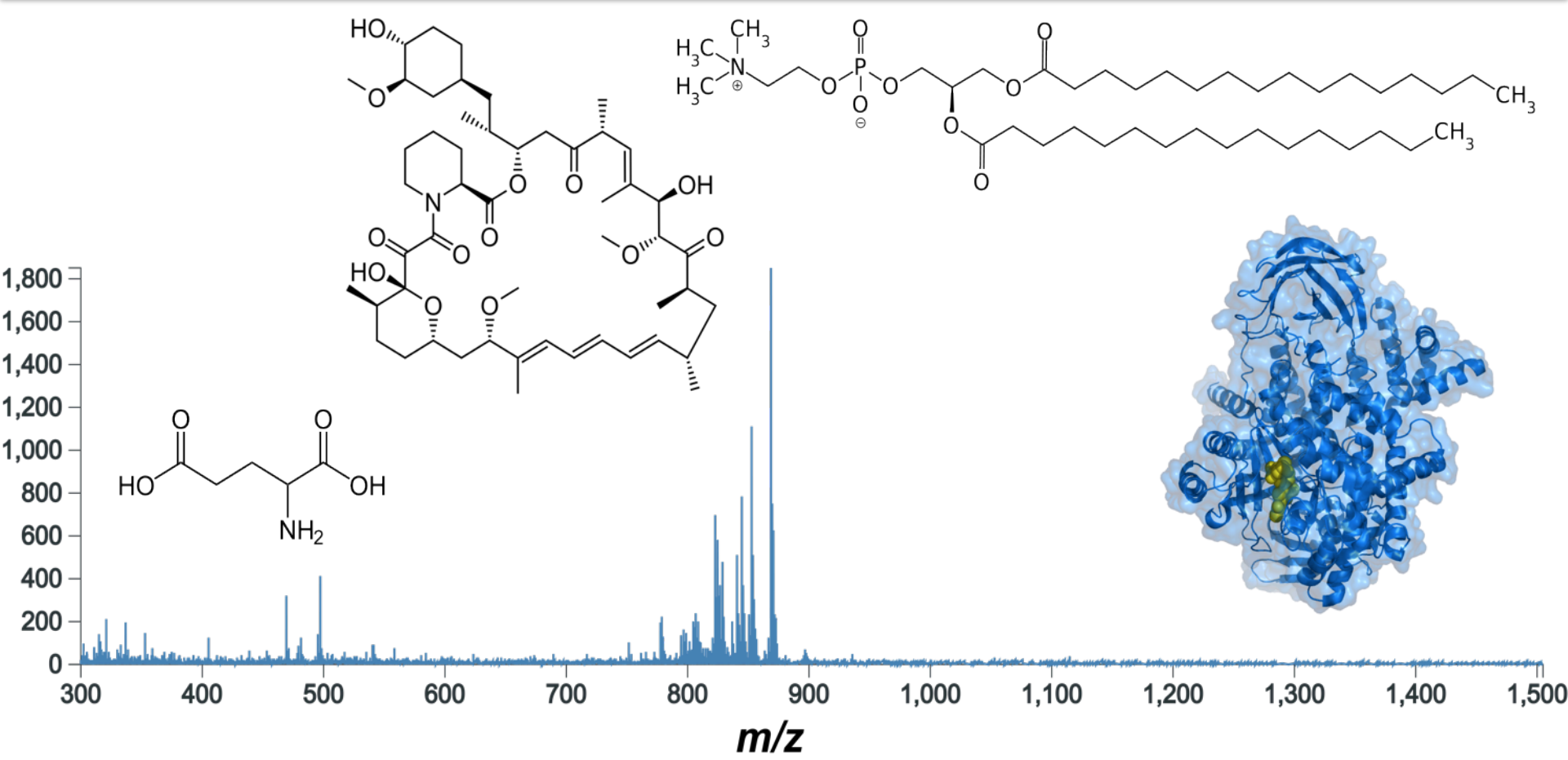
Raster scanning a laser across a brain tissue sample and laser ablation effects

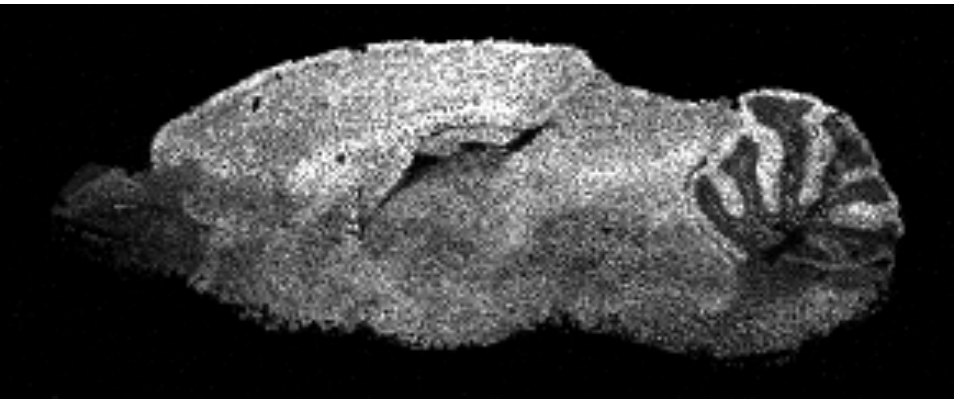


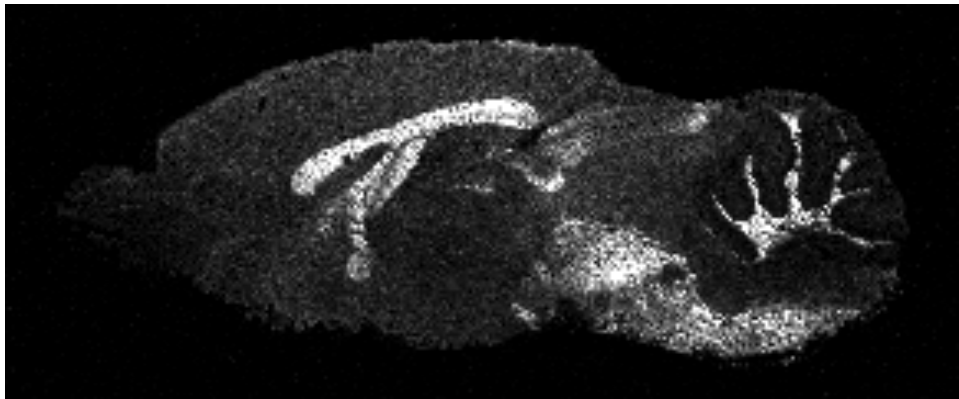


Chemical images provide complex measurements of the spatial distribution of molecules

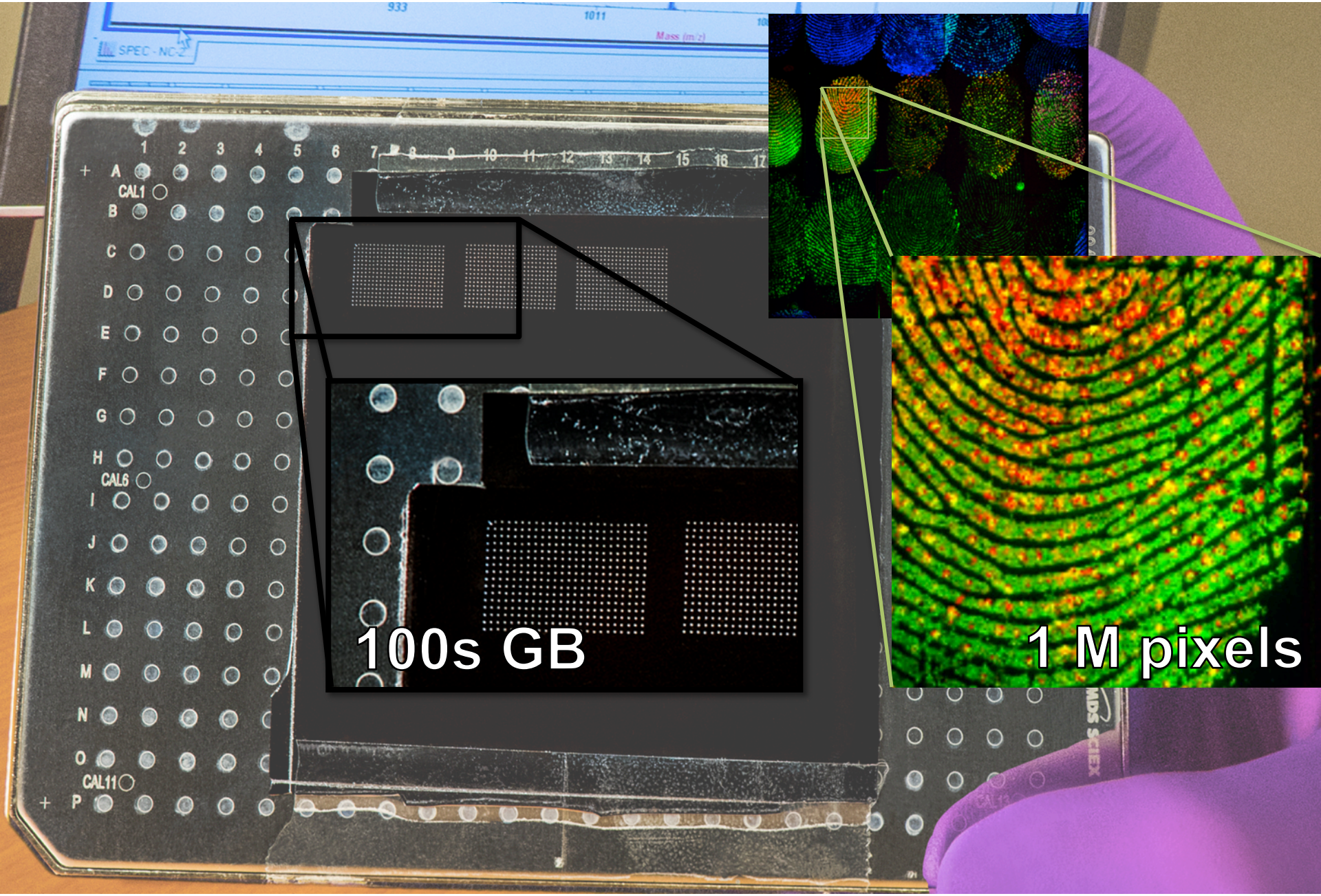
MSI sees 1° and 2° metabolites, lipids, and proteins



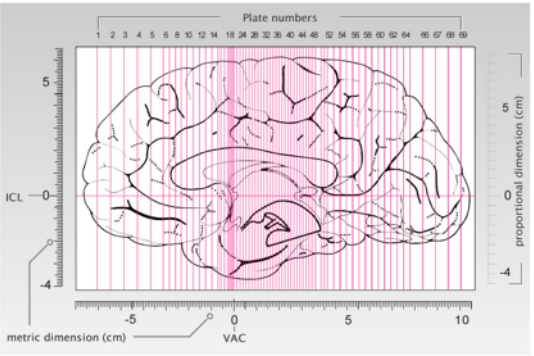


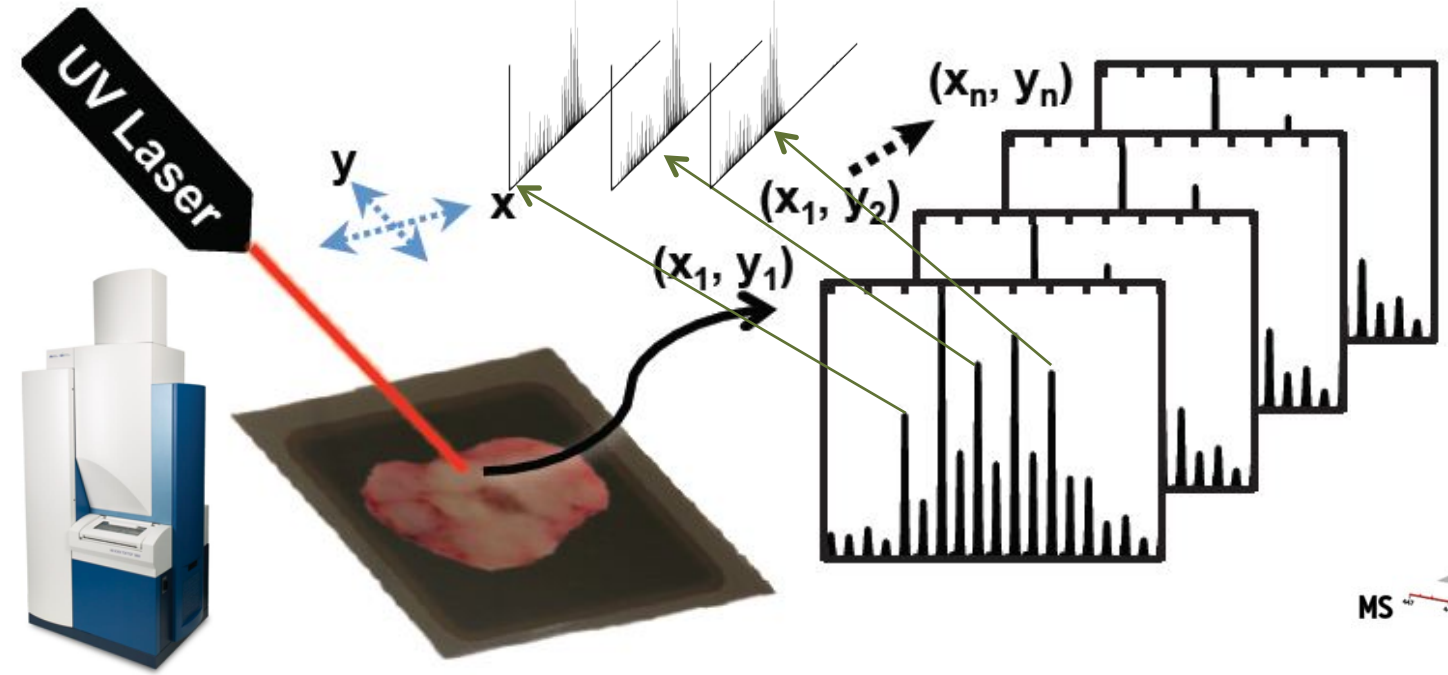


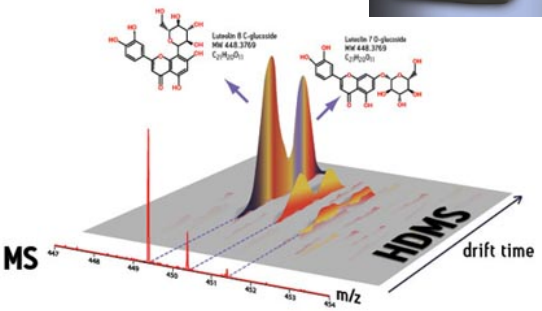
MSI files can be fairly large

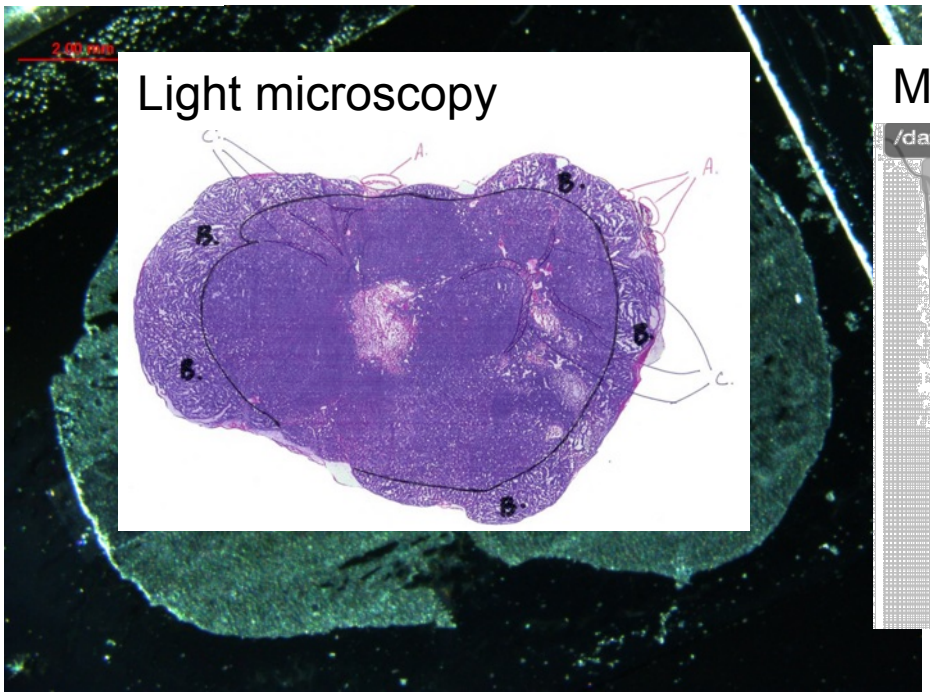


Frankenstein's dilemma



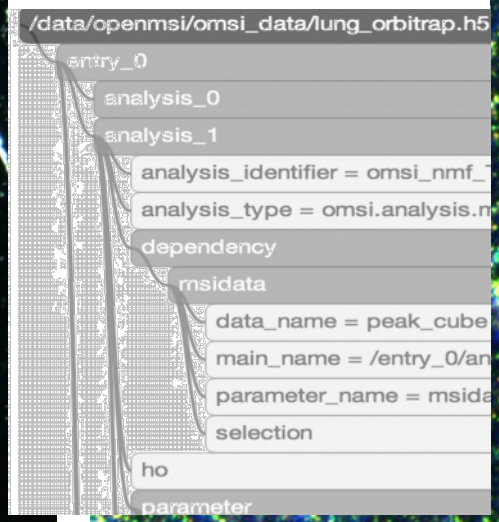




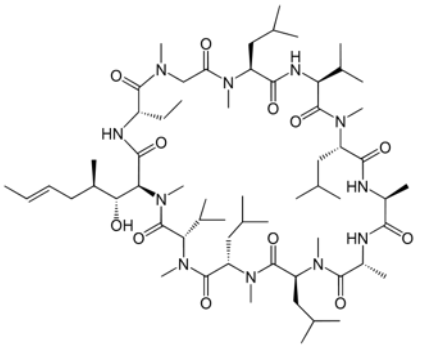


Light microscopy

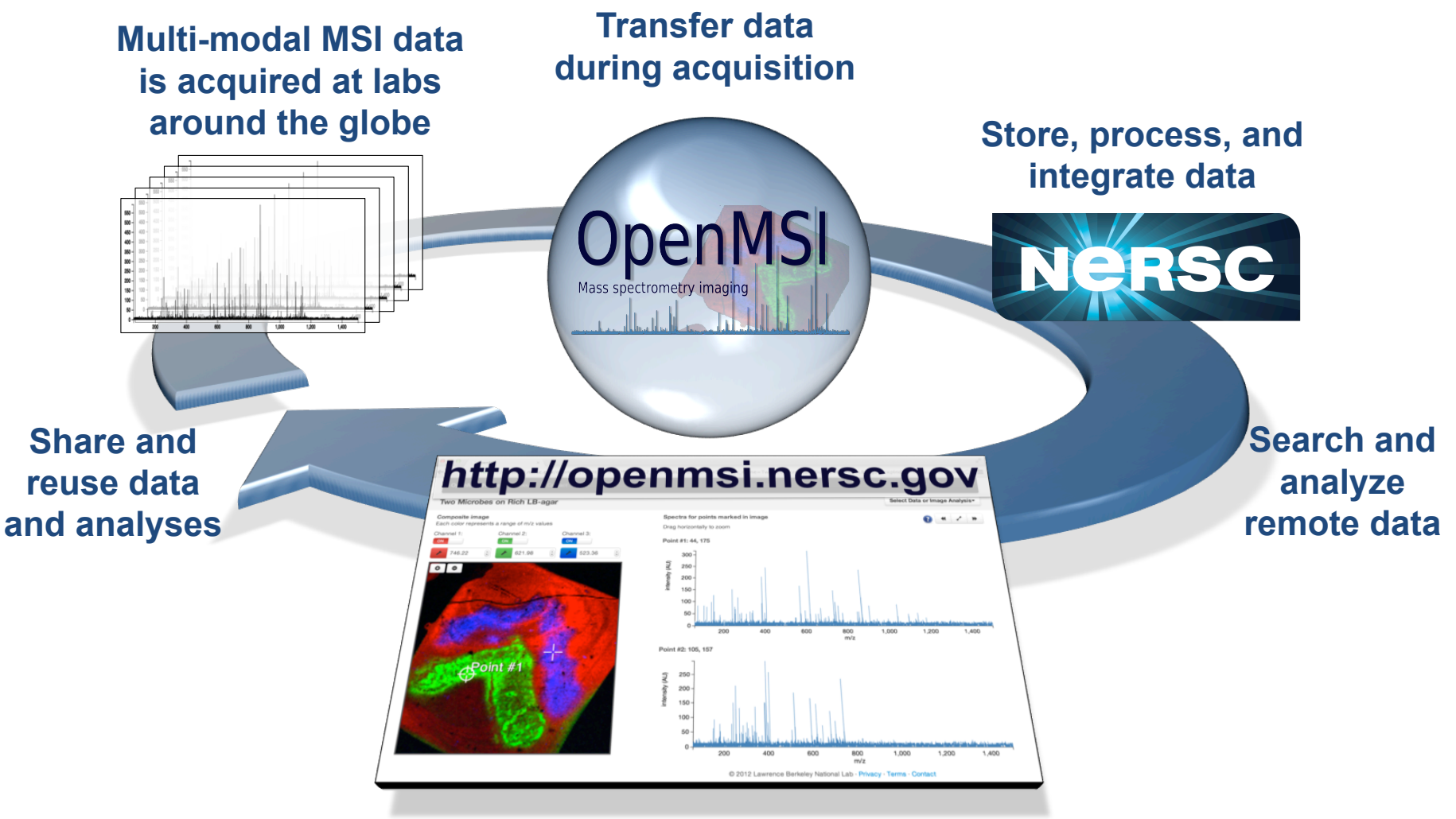
Metadata



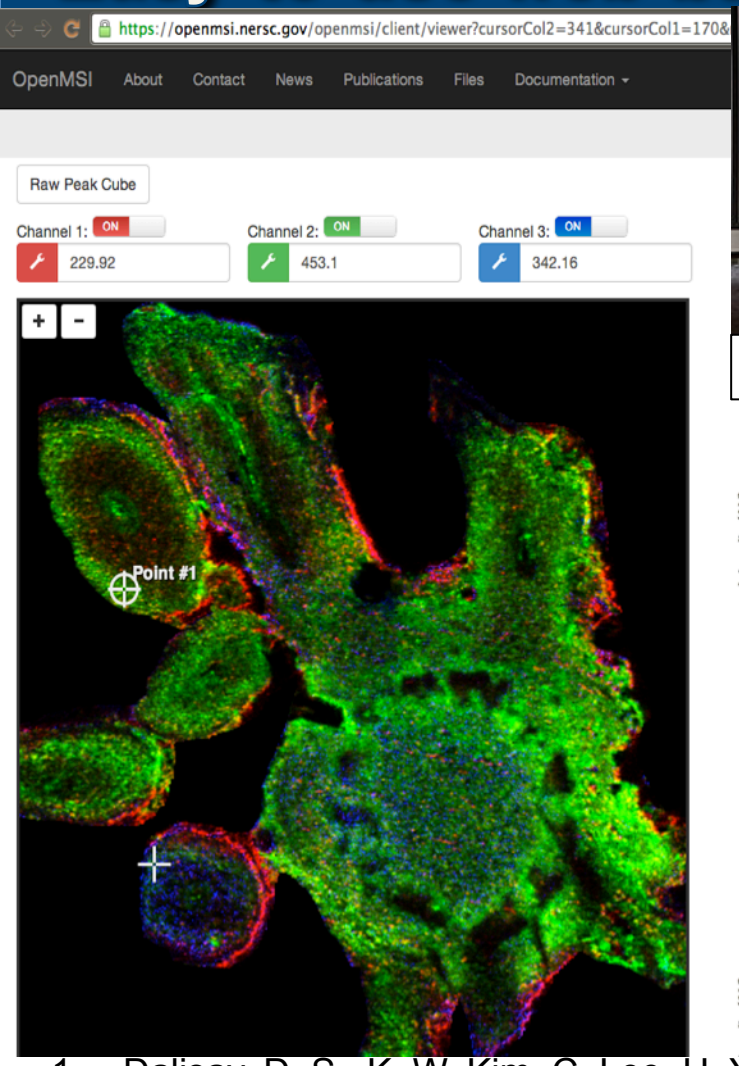
Compound IDs

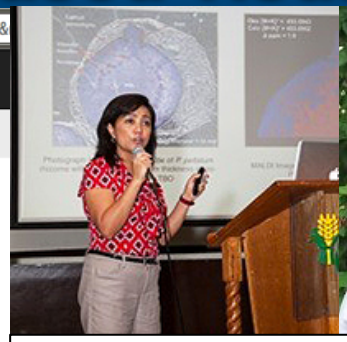


This can't be done on your desktop or laptop



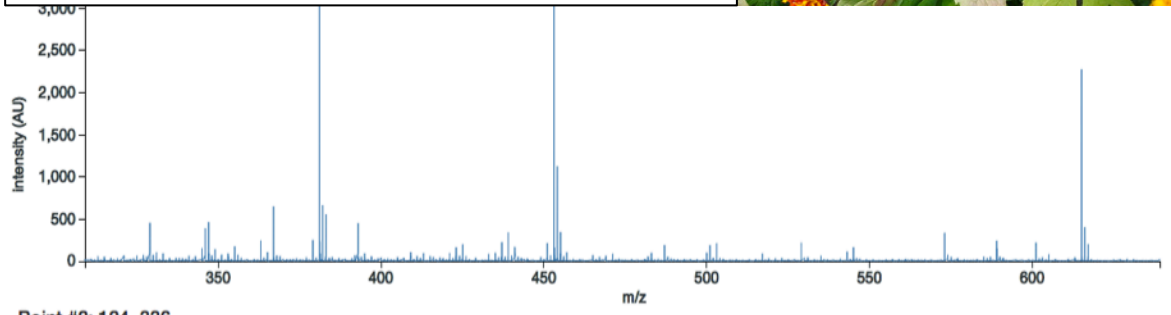
Easy-to-use web-based visualization





Podophyllum hexandrum

Doralyn Dalisay & Norman Lewis (WSU)

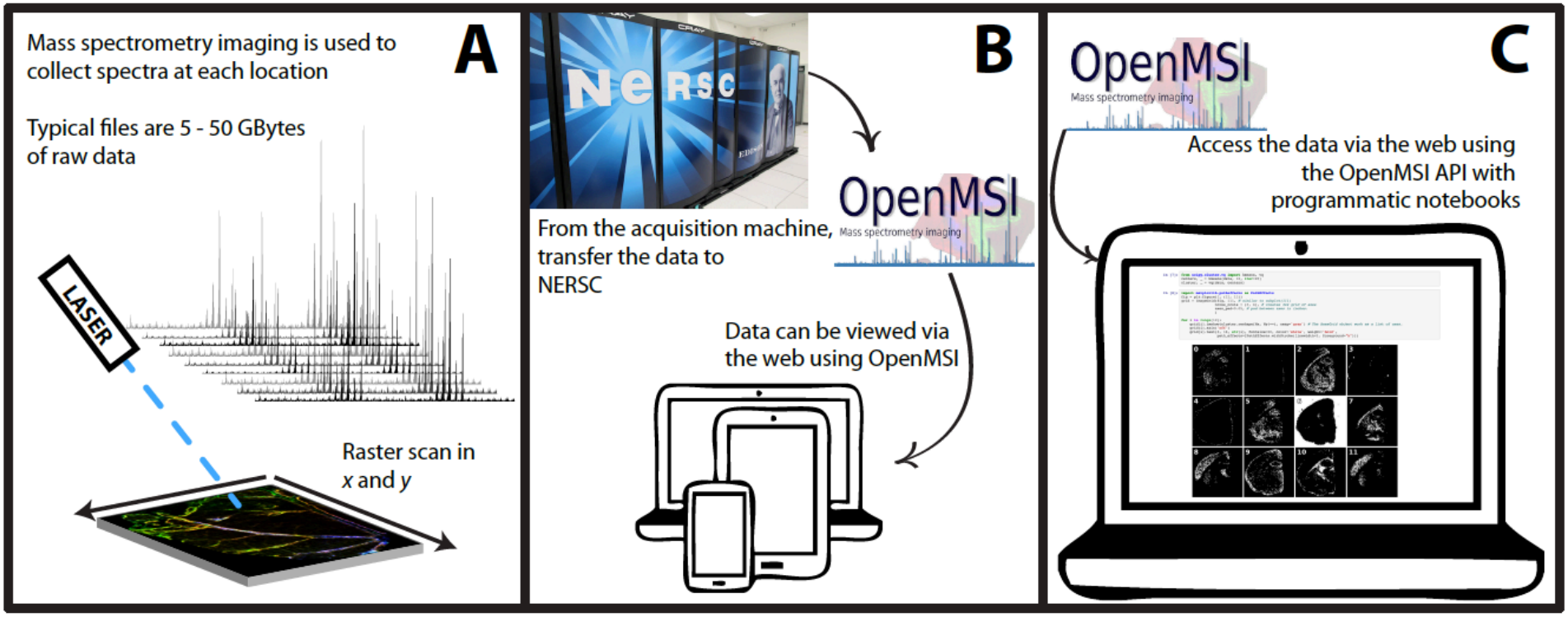


Point #2: 124, 336

1. Dalisay, D. S., K. W. Kim, C. Lee, H. Yang, O. Rübel, B. P. Bowen, L. B. Davin, and N. G. Lewis. "Secoisolariciresinol diglucoside formation and cyanogenic glucosides in flax seed: MALDI mass spectrometry imaging." Journal of natural products (2015)
2. Marques, Joaquim V., Doralyn S. Dalisay, Hong Yang, Choonseok Lee, Laurence B. Davin, and Norman G. Lewis. "A multi-omics strategy resolves the elusive nature of alkaloids in Podophyllum species." Molecular BioSystems 10, no. 11 (2014): 2838-2849.

Sharable, custom analyses of MSI data via the web

- The analysis only needs to implement the basic data patterns: `qmz`, `qspectrum` and `qslice`
- Fast deployment of new analyses to application scientists during development
- Analysis is shared independently of data



http://tinyurl.com/openmsi-nb1
http://tinyurl.com/openmsi-nb2

Acknowledgements

Metabolite Atlas Team:

Steven Silvester

Jimmy Touma

Curt Fischer

Lisa Gerhardt

Yushu Yao

Oliver Ruebel

Terence Sun

Yony Wang

Richard Baran

Trent Northen

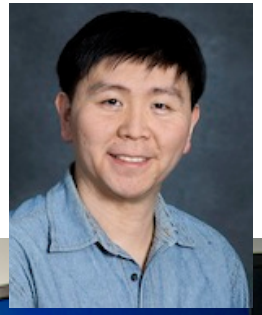
OpenMSI Team:

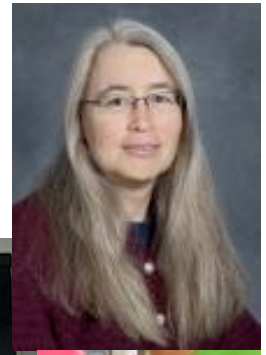
Oliver Ruebel

Shreyas Cholia

Annette Greiner

Many More!





JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

jbei Joint BioEnergy Institute

Biological and Environmental Research Program

ENIGMA