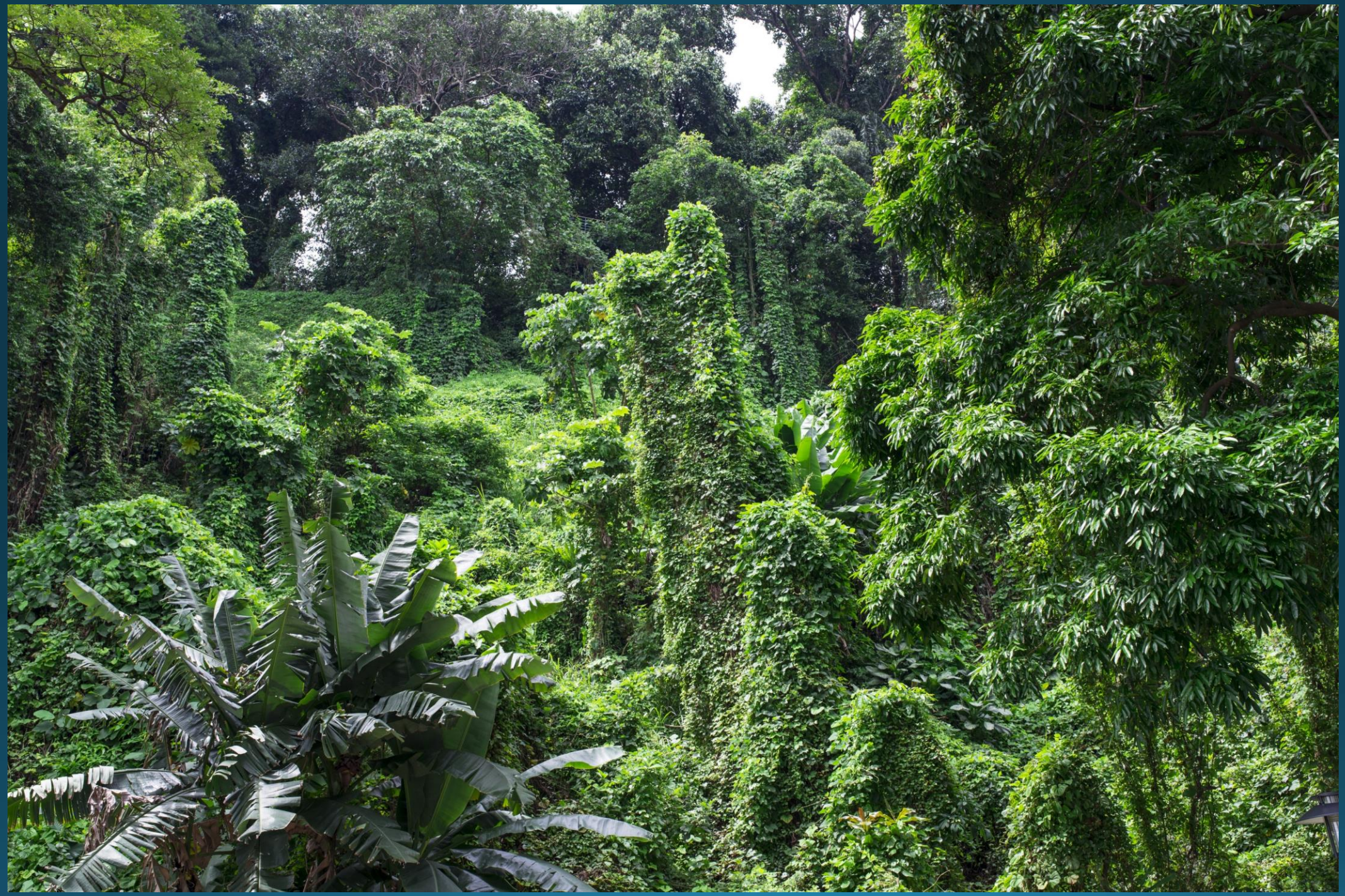
Cheaper by the Million:

Expanding Untargeted Metabolomics into the Theoretical Realm

Anthony Mills

Coley/Kursar Laboratory

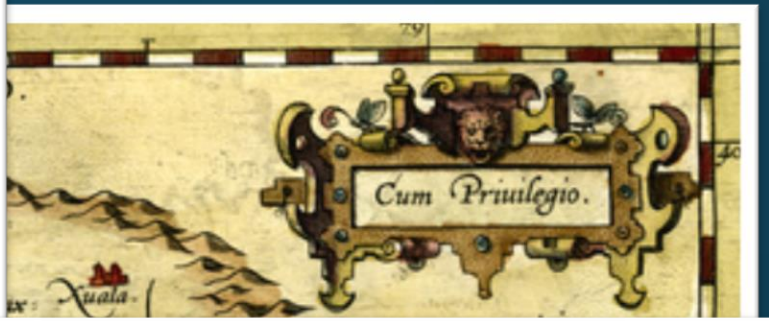
University of Utah Department of Biology

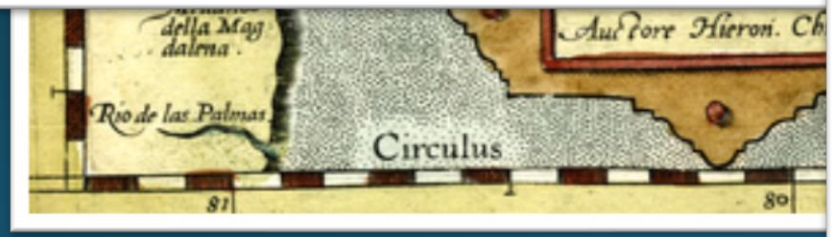


Coley/Kursar Lab Question:

Why is the rainforest green?

Juan Ponce de León


Cum Priuilegio.


della Mag
dalena
Rio de las Palmas
Circulus
81
80

Hippomane mancinella

“La manzanilla de la muerte”


MANZANILLA FRUITS
POISONOUS
GIFTIG
VENENOSO
MANCHINEEL TREES


Euphorbiaceae
Manchineel
Hippomane mancinella
DO NOT TOUCH!!!
This tree is very TOXIC ! The fruit is poisonous and the sap from leaves and stems can produce painful blisters. Do not stand under this tree during rain as the water on the leaves can pick up the tozin and drip it on the skin.
1989113 West Indies, Trop. America, Florida

Who’s the real enemy?



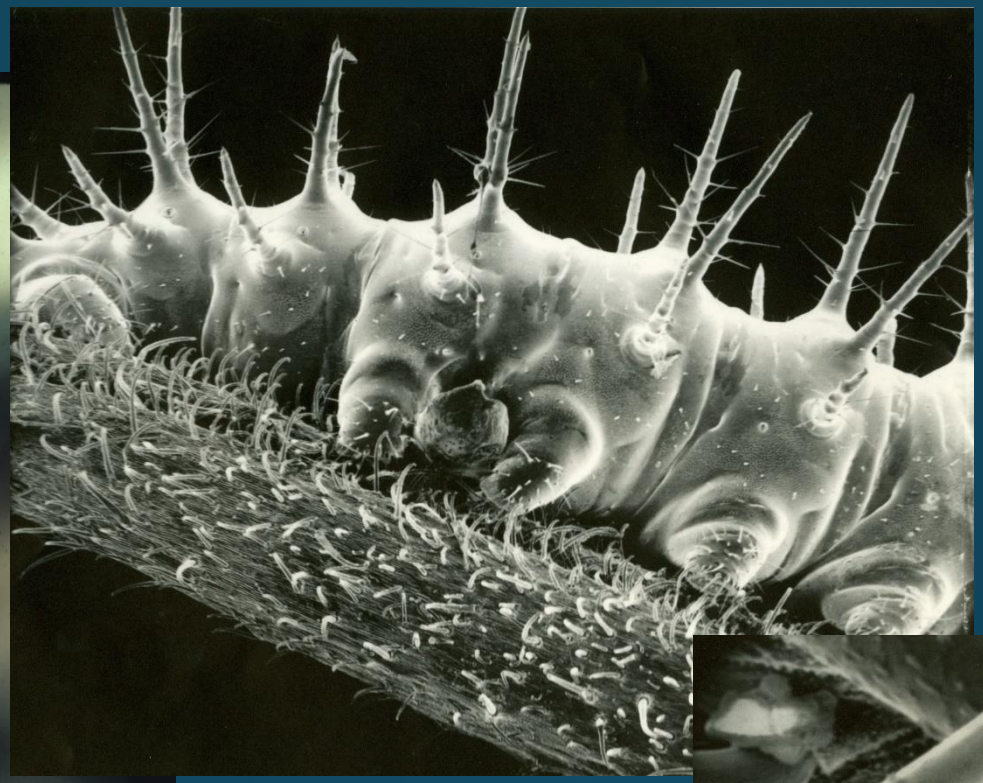


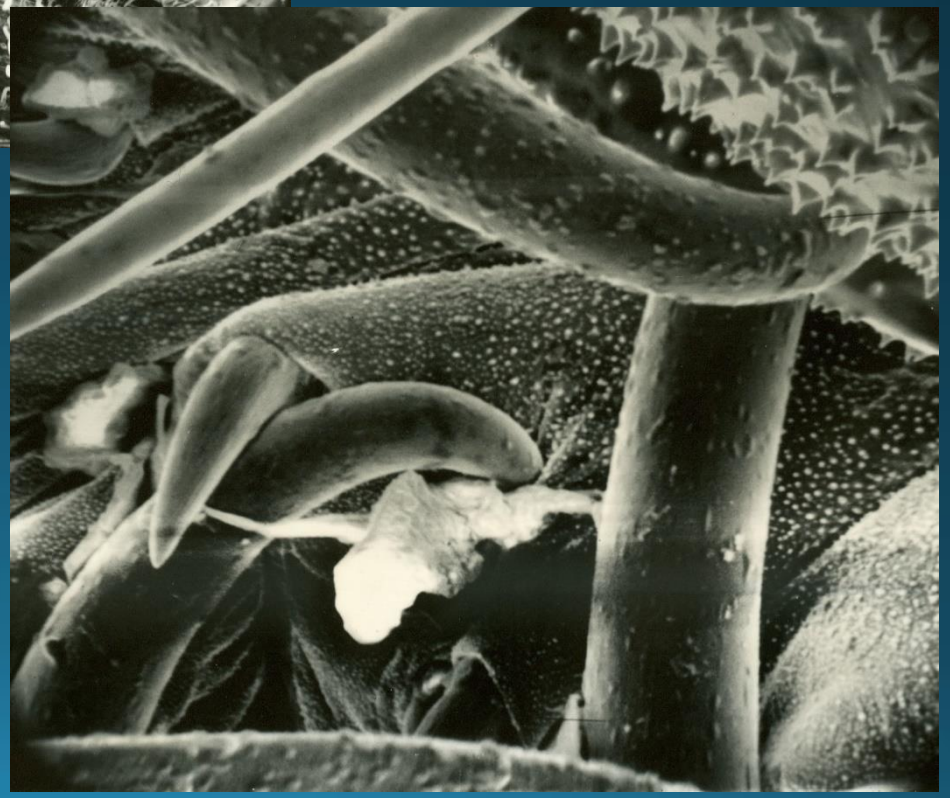




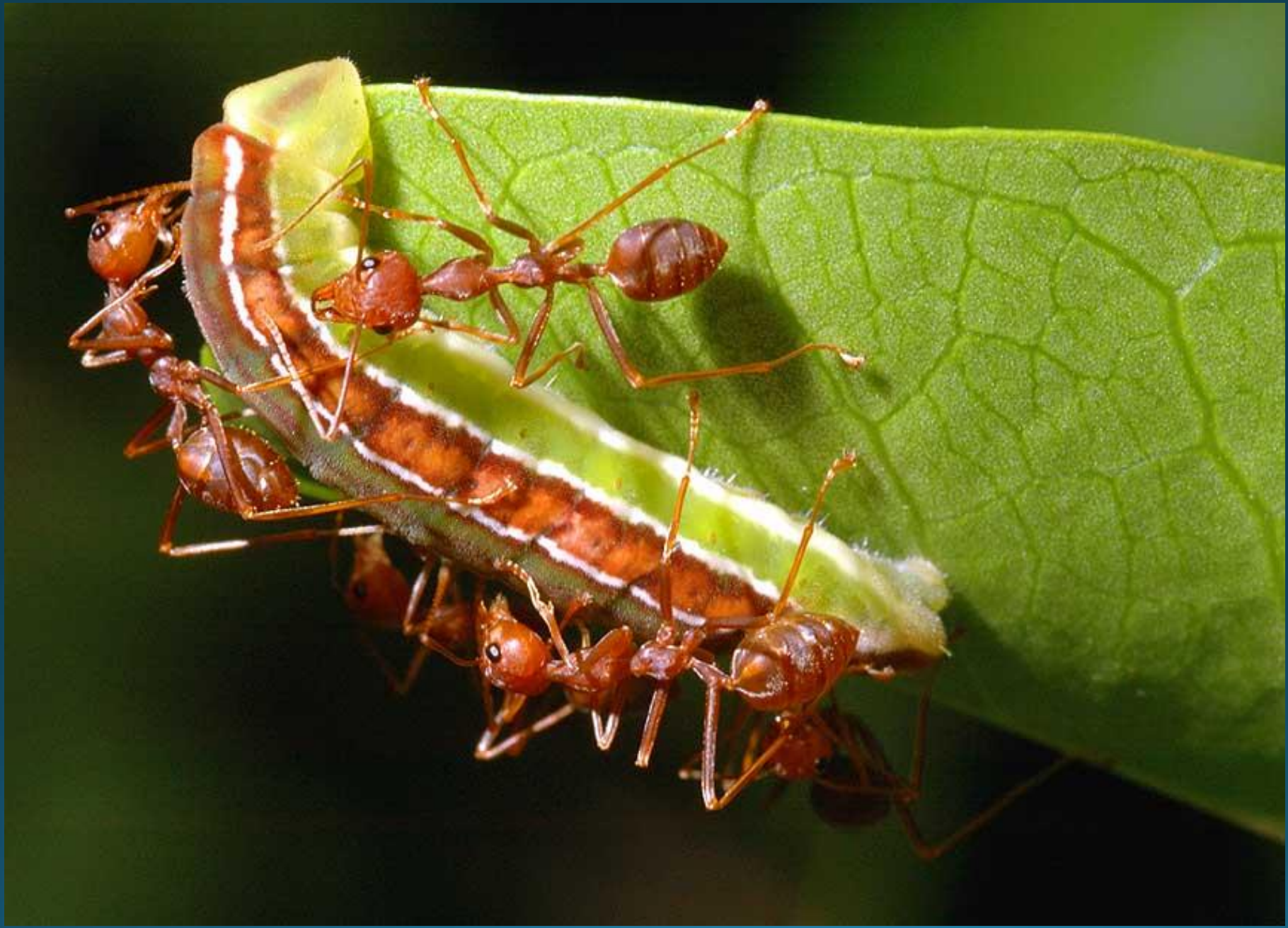


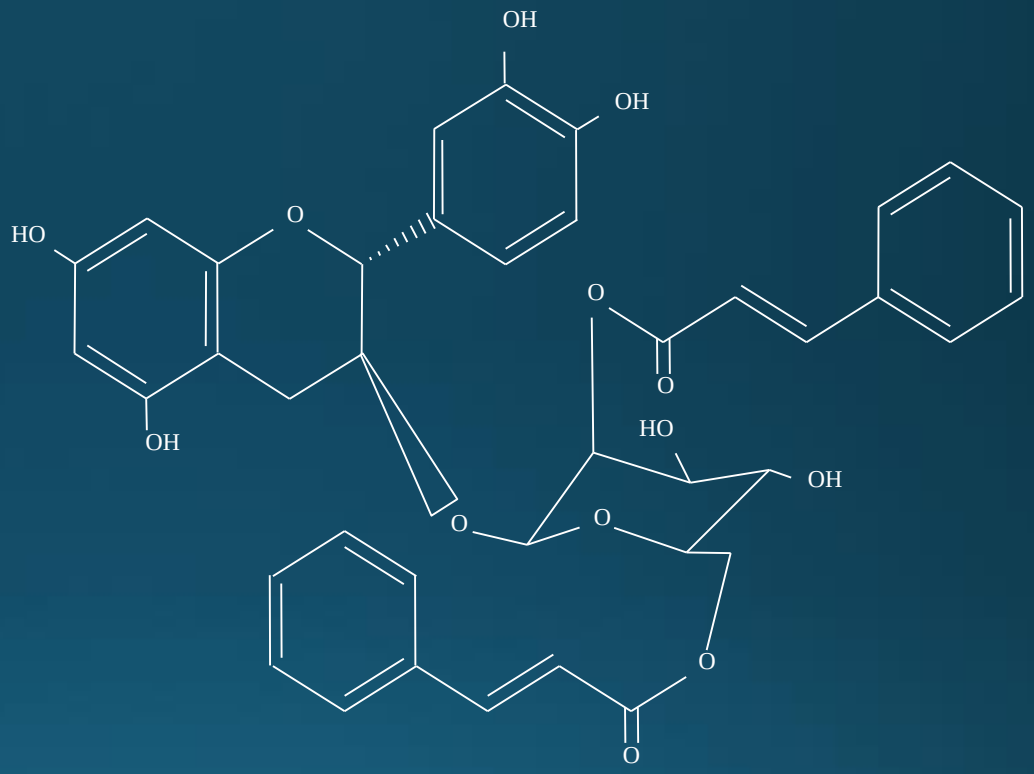

≠



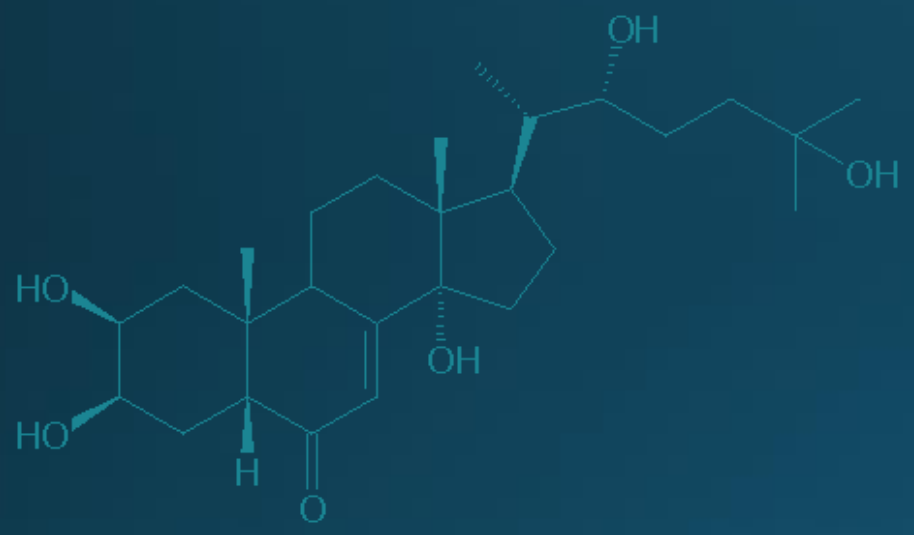


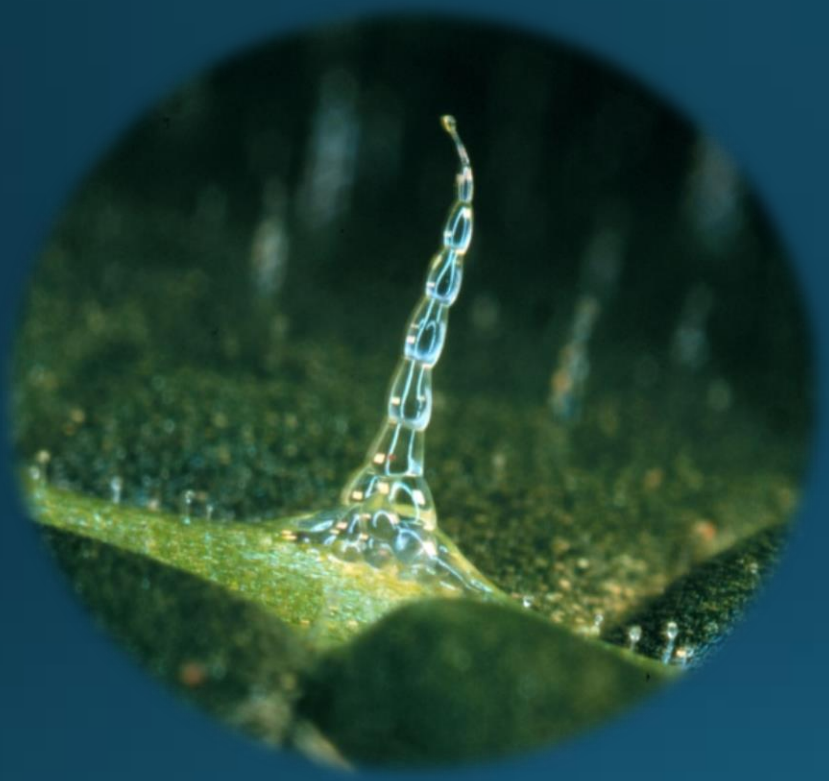

© Jeff Cremer / @JCremerPhoto

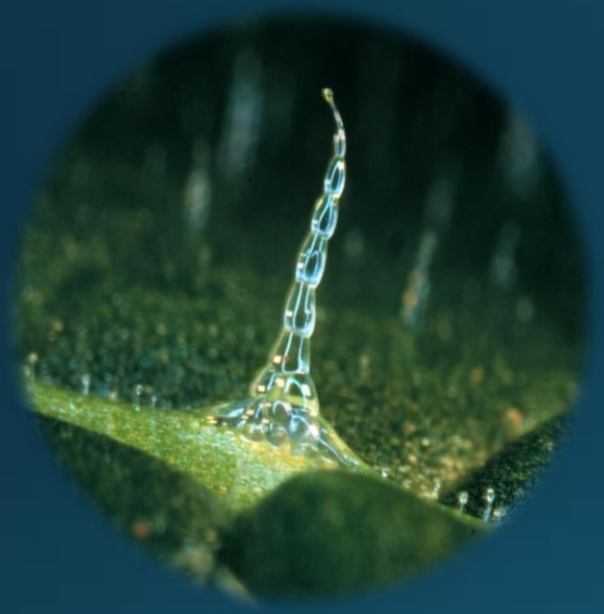



OH
OH
HO
O
O
O
OH
HO
OH
O
O
O
O

Ecdysone







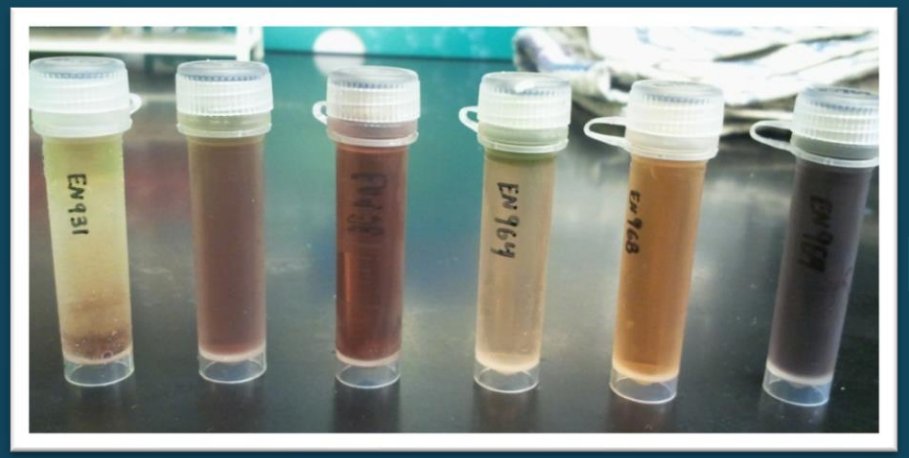
Plant-Herbivore Interactions

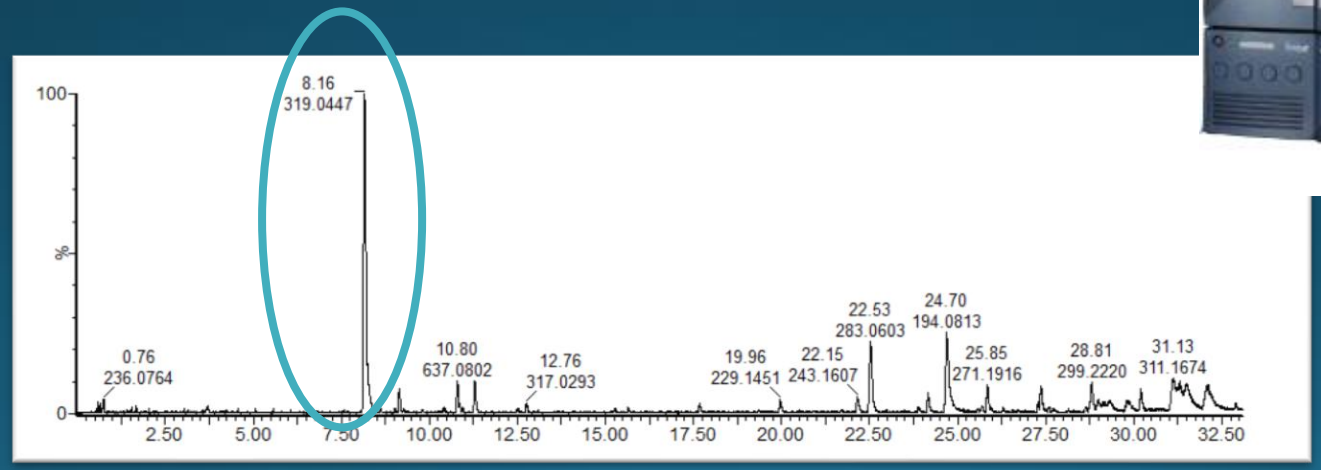
Inga

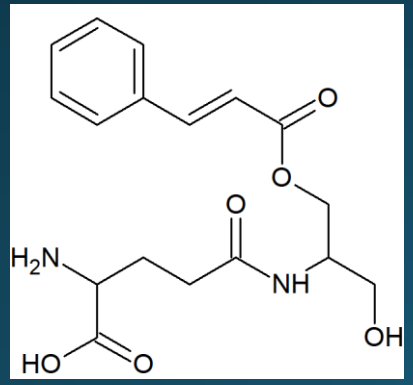
- Neotropical genus
- Legume (Fabaceae)
- What's in them?
  - 3,600 Compounds
    - Only 4% known
  - Ecologically important
  - Central goal of lab

Procedure

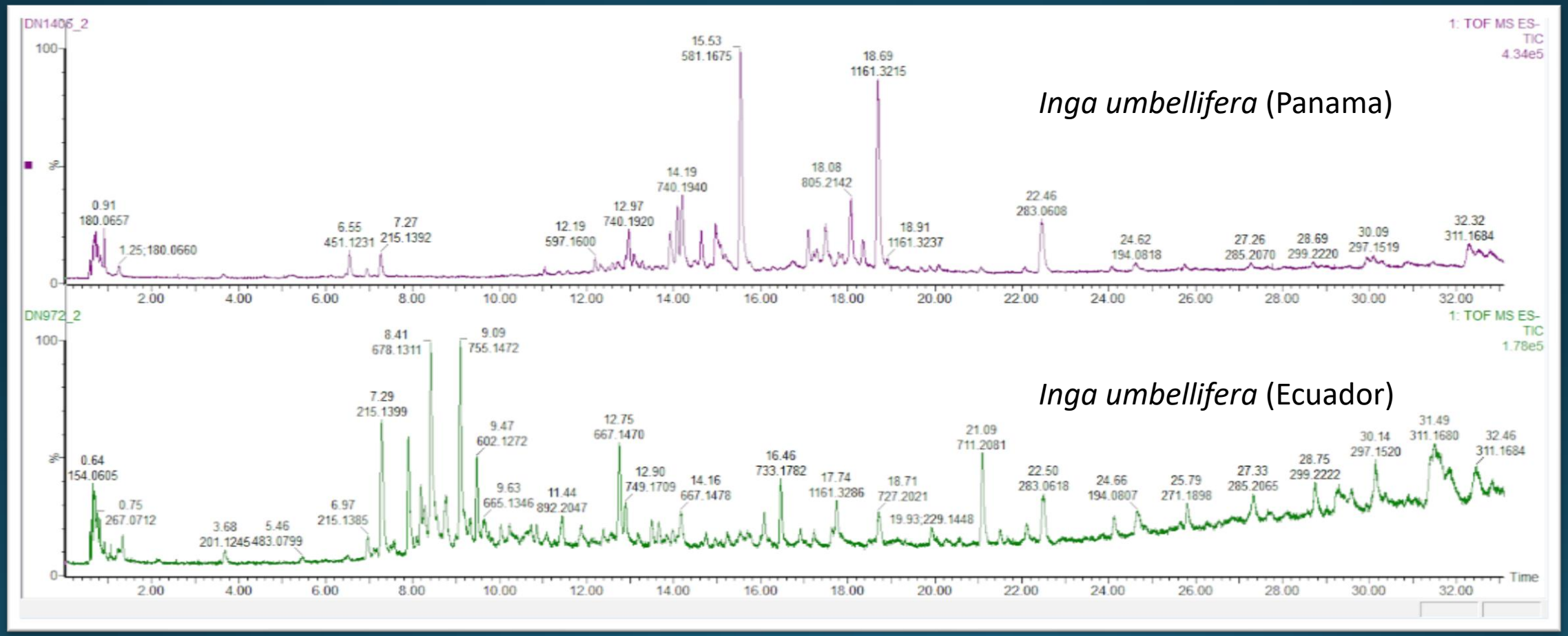

14


EN931
EN964


100
%
0
8.16
319.0447
0.76
236.0764
10.80
637.0802
12.76
317.0293
19.96
229.1451
22.15
243.1607
22.53
283.0603
24.70
194.0813
25.85
271.1916
28.81
299.2220
31.13
311.1674
2.50
5.00
7.50
10.00
12.50
15.00
17.50
20.00
22.50
25.00
27.50
30.00
32.50
time


O
O
O
H_2N
NH
OH
HO
O

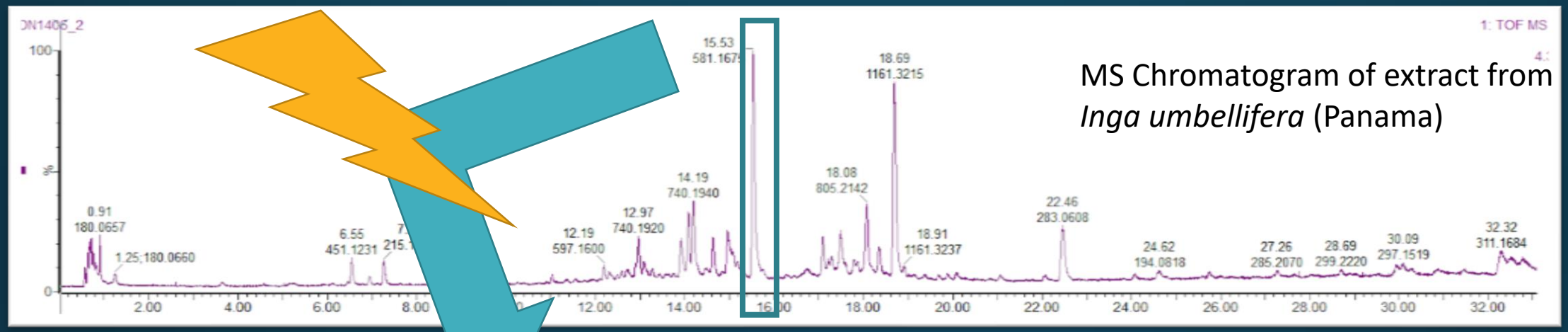
Chemotypes


DN1406_2
1: TOF MS ES-
TIC
4.34e5
Inga umbellifera (Panama)
15.53
581.1675
18.69
1161.3215
18.08
805.2142
14.19
740.1940
12.97
740.1920
0.91
180.0657
1.25;180.0660
6.55
451.1231
7.27
215.1392
12.19
597.1600
18.91
1161.3237
22.46
283.0608
24.62
194.0818
27.26
285.2070
28.69
299.2220
30.09
297.1519
32.32
311.1684
DN972_2
1: TOF MS ES-
TIC
1.78e5
Inga umbellifera (Ecuador)
8.41
678.1311
9.09
755.1472
7.29
215.1399
9.47
602.1272
12.75
667.1470
0.64
154.0605
0.75
267.0712
3.68
201.1245
5.46
483.0799
6.97
215.1385
9.63
665.1346
11.44
892.2047
12.90
749.1709
14.16
667.1478
16.46
733.1782
17.74
1161.3286
18.71
727.2021
19.93;229.1448
21.09
711.2081
22.50
283.0618
24.66
194.0807
25.79
271.1898
27.33
285.2065
28.75
299.2222
30.14
297.1520
31.49
311.1680
32.46
311.1684
Time

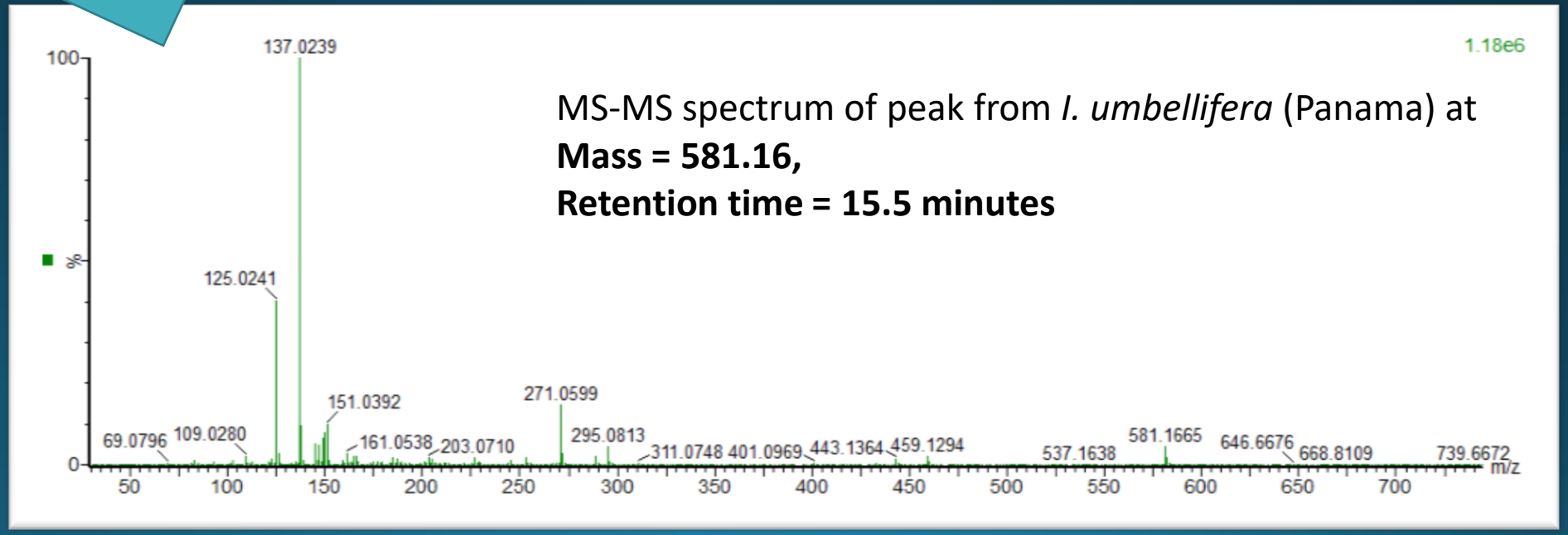
Goal: Improve Identification

- Purification and NMR are costly and time+labor intensive
- Untargeted metabolomics still a developing field
- High-Res Mass is starting point
- MS-MS gives structural information

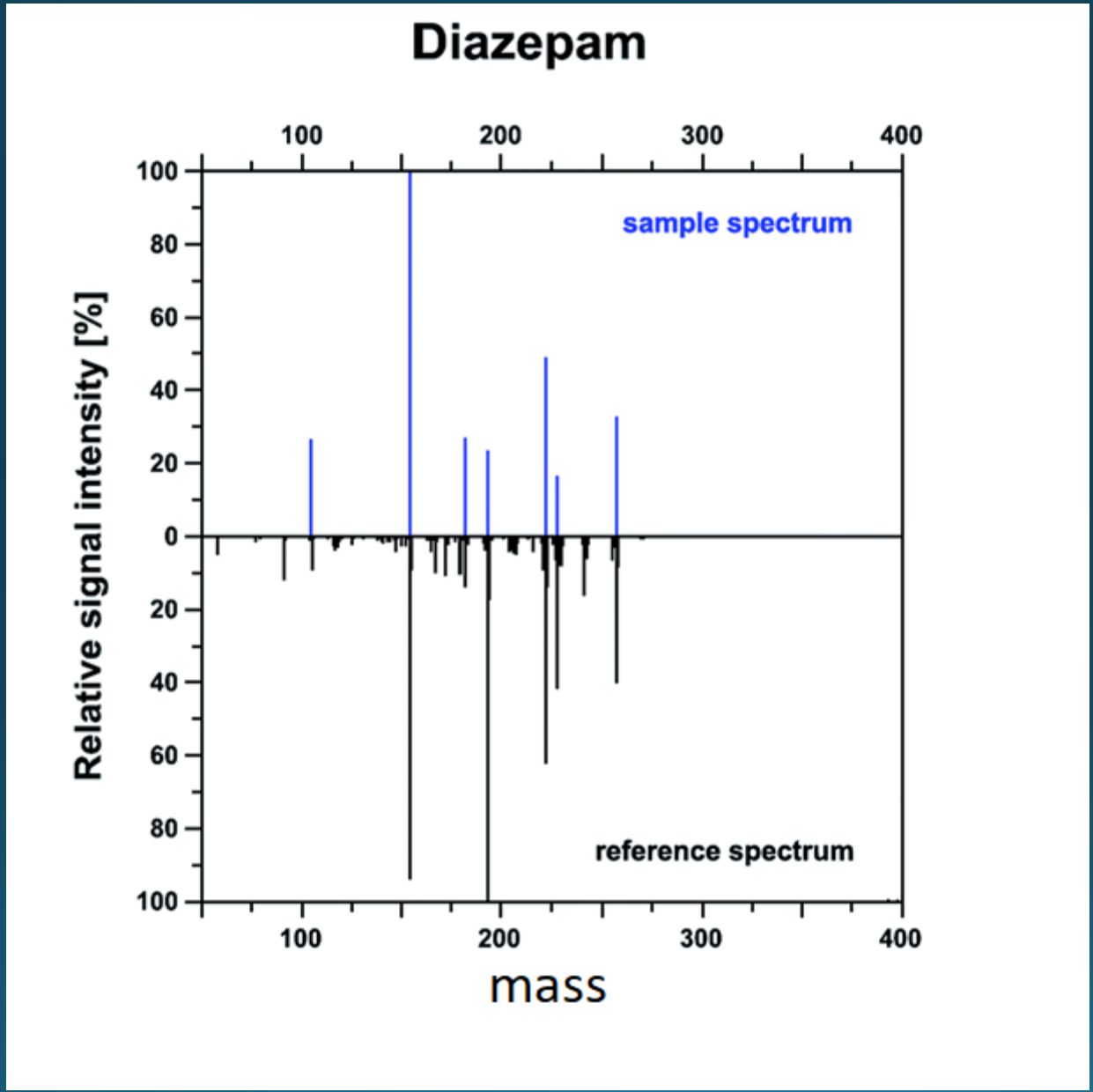
Tandem Mass Spectrometry



MS Chromatogram of extract from *Inga umbellifera* (Panama)



MS-MS spectrum of peak from *I. umbellifera* (Panama) at
Mass = 581.16,
Retention time = 15.5 minutes


Diazepam
sample spectrum
reference spectrum
Relative signal intensity [%]
mass
100
200
300
400
100
80
60
40
20
0
20
40
60
80
100

Universal Natural Products Database

- NP Databases do exist
- Few if any MS-MS spectra
- Structures present

In-Silico fragmentation

CFM-ID | Utilities | Help | Data | Publications | Contact Us

CFM-ID
Competitive Fragmentation Modeling for Metabolite Identification

Spectra Prediction

Predicts the spectra for a given input molecule. Spectra are computed for low (10V), medium (20V) and high (40V) collision energy levels and are represented by a list of 'mass intensity' pairs, each corresponding to a peak in the spectra.

Parent Compound Structure
InChI or SMILES format

Enter an InChI or SMILES string

InChI strings need to start with "InChI=" and are not expected to have any charge - an additional H+ will be added. Maximum compound size is 200 atoms. Load an InChI example, SMILES example, or another SMILES example.

Spectra Type ESI

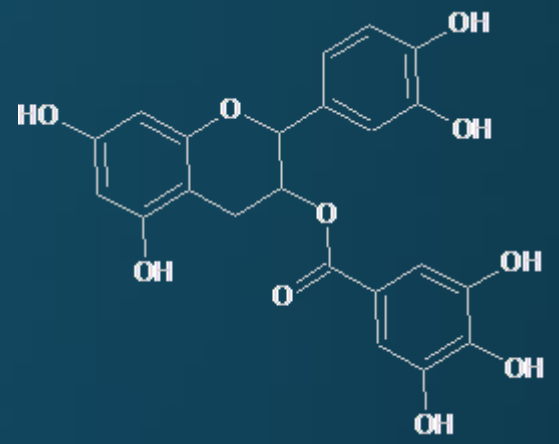
Ion Mode Positive

Adduct Type [M+H]+

Reset Submit

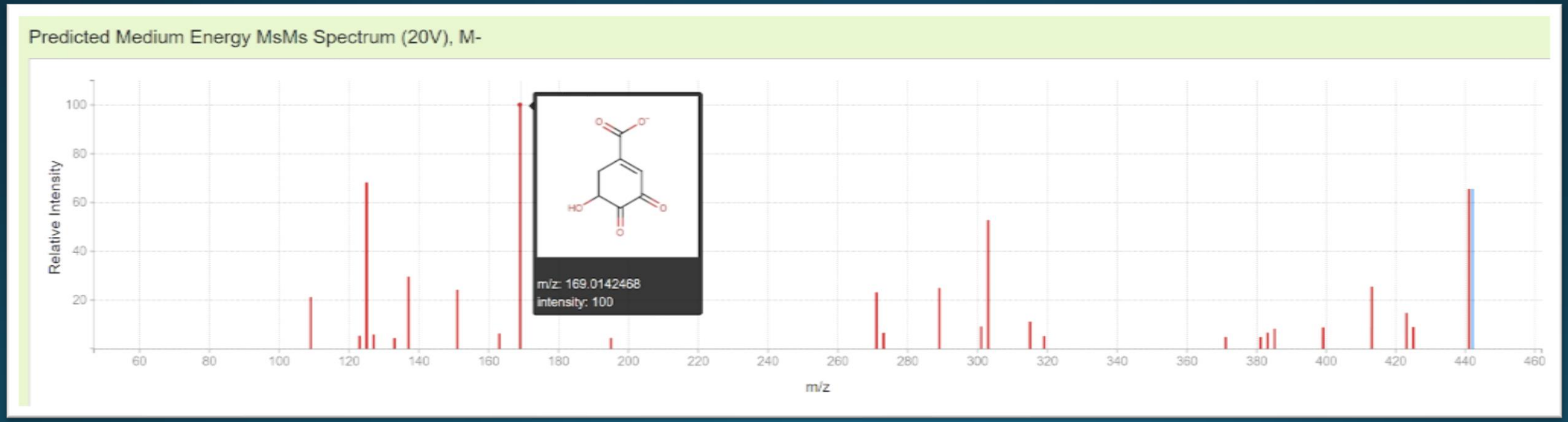
If you wish to run multiple jobs, input larger query molecules, or customize the computation parameters, you can freely download the source code here: http://sourceforge.net/projects/cfm-id.

In-Silico fragmentation



Epicatechin Gallate:

O=C(O[C@@H]2Cc3c(O[C@@H]2c1ccc(O)c(O)c1)cc(O)cc3O)c4cc(O)c(O)c(O)c4



~230,000 Compounds

CHPC

Ember Cluster

- 144 Dual Socket-Six Core Nodes (1728 total cores)
- 2.8 GHz Intel Xeon (Westmere X5660) processors
- RAM:
  - 24 Gbytes memory per node (2 Gbytes per processor core) on most nodes

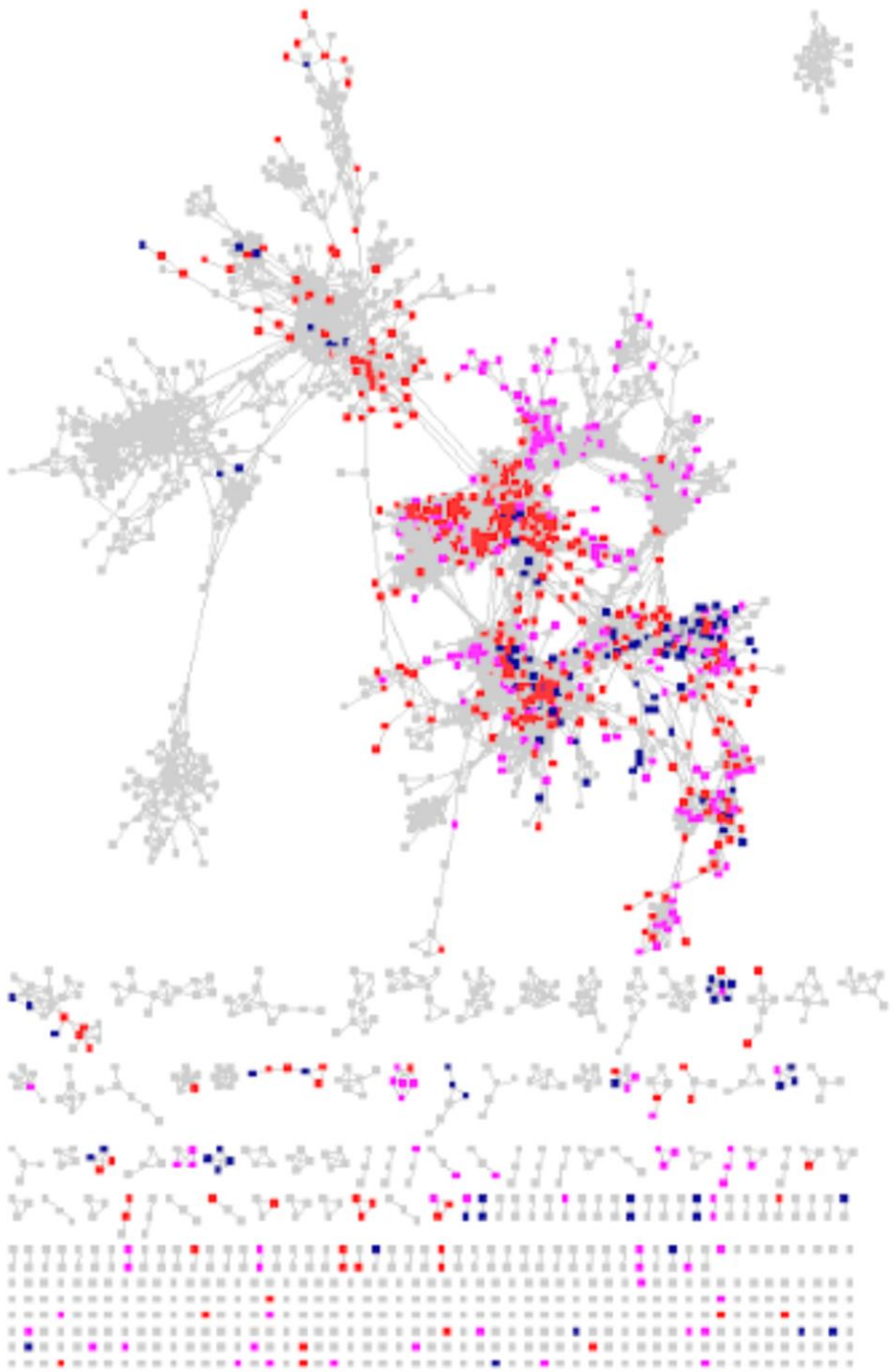
Open Science Grid

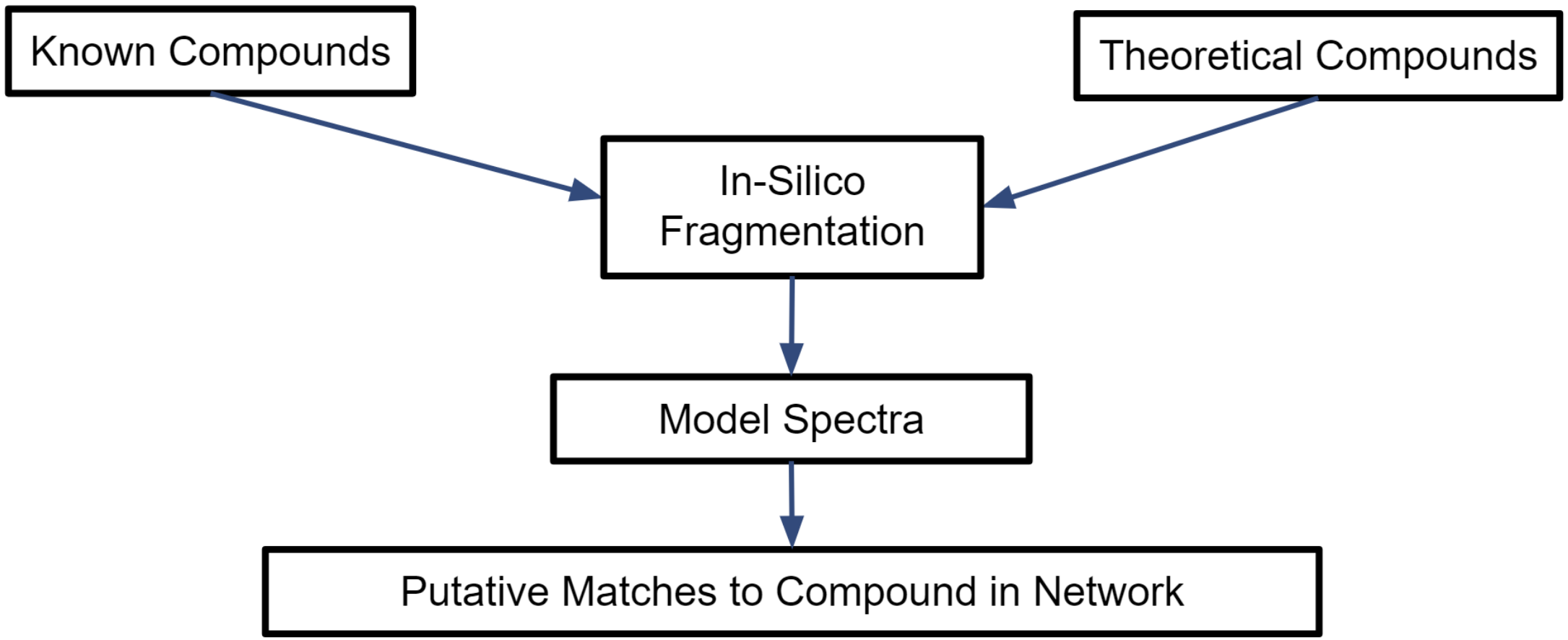
Comparison of HTC to HPC

230,737 molecules total	Job Size	Run Rate	Free Nodes	Time to Completion
Ember Cluster CHPC	462 files containing 500 molecules each	69 molecules / hour	10-20	334 hours (based on 10 free nodes)
OSG	4,614 files containing 50 molecules each	4 molecules / hour	Thousands	12 hours

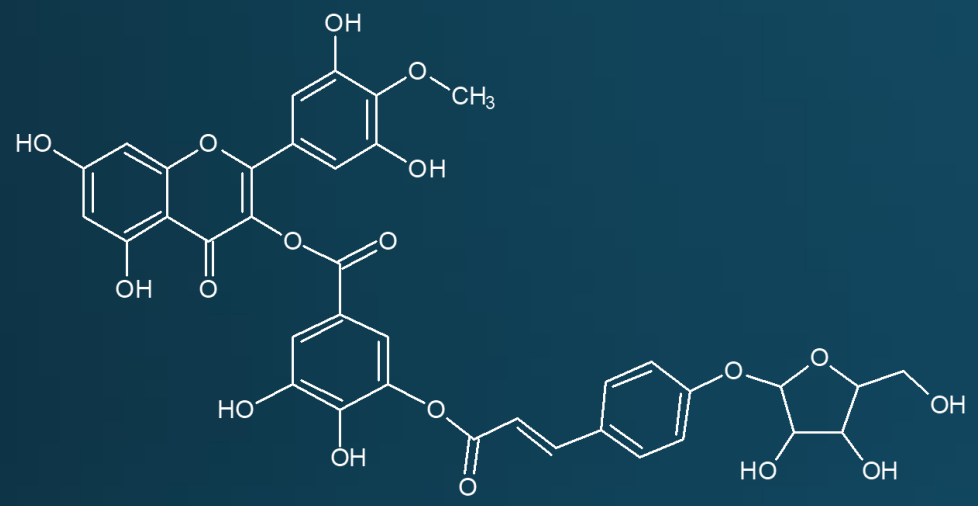
Results:

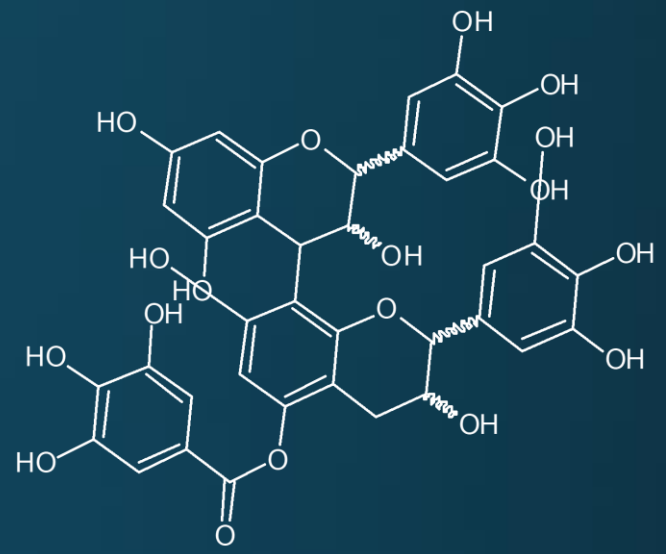
- **Blue** = match to Library MS-MS spectrum
- **Red** = match to *in-silico* MS-MS spectrum
- **Magenta** = match to mass only

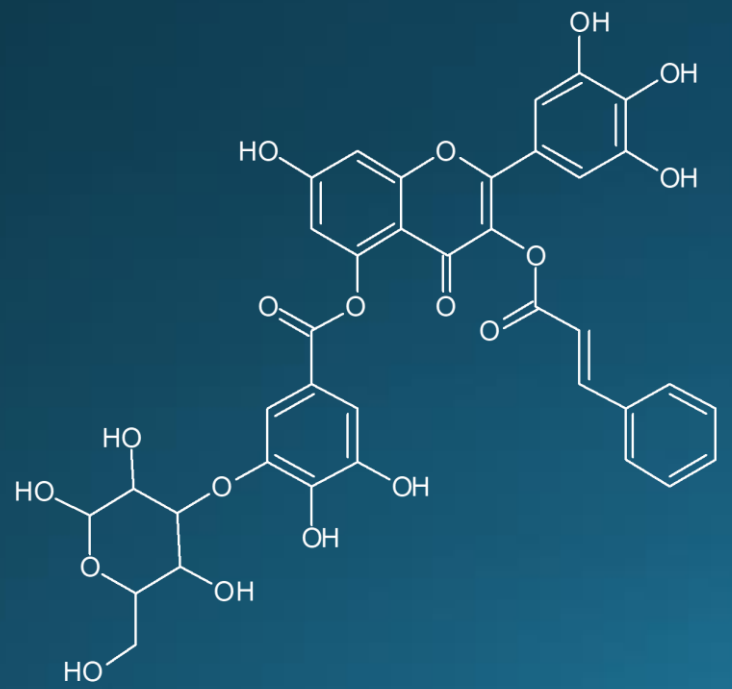


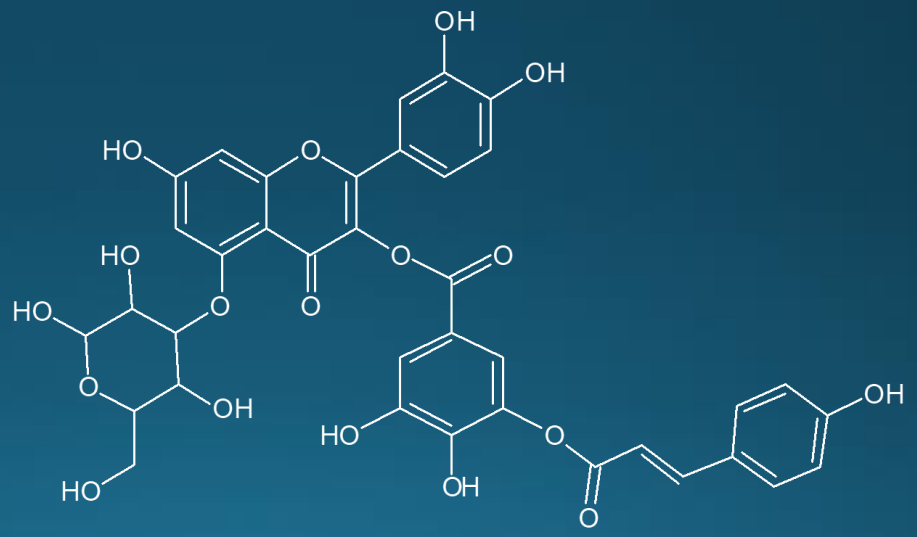

Known Compounds
Theoretical Compounds
In-Silico Fragmentation
Model Spectra
Putative Matches to Compound in Network

Combinatorial Chemistry

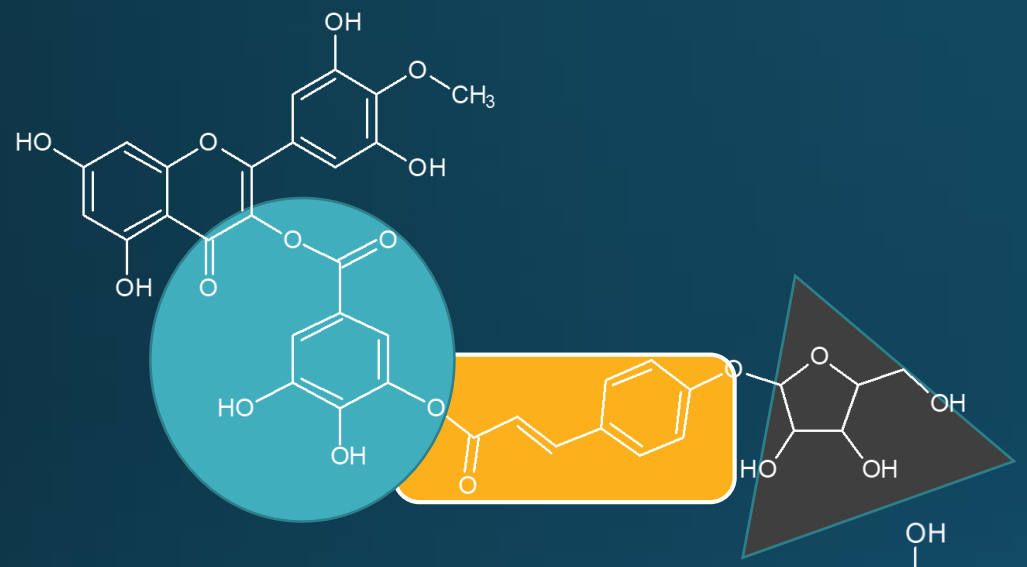


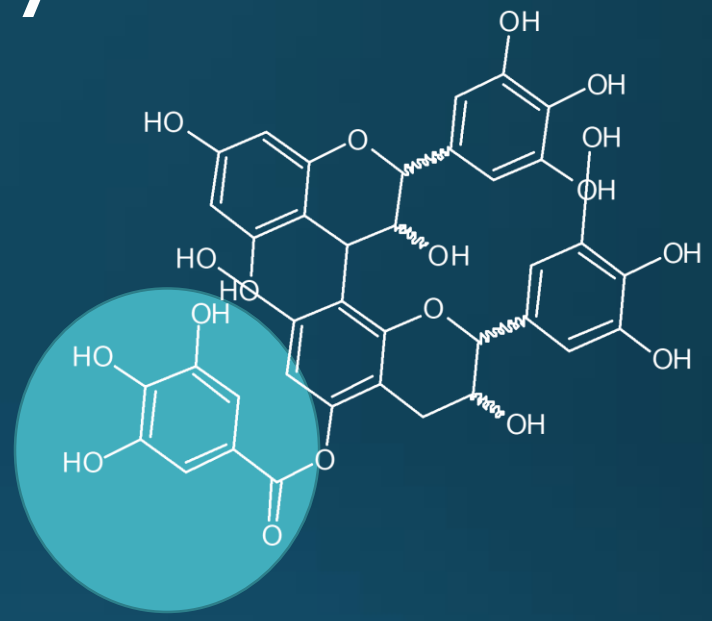


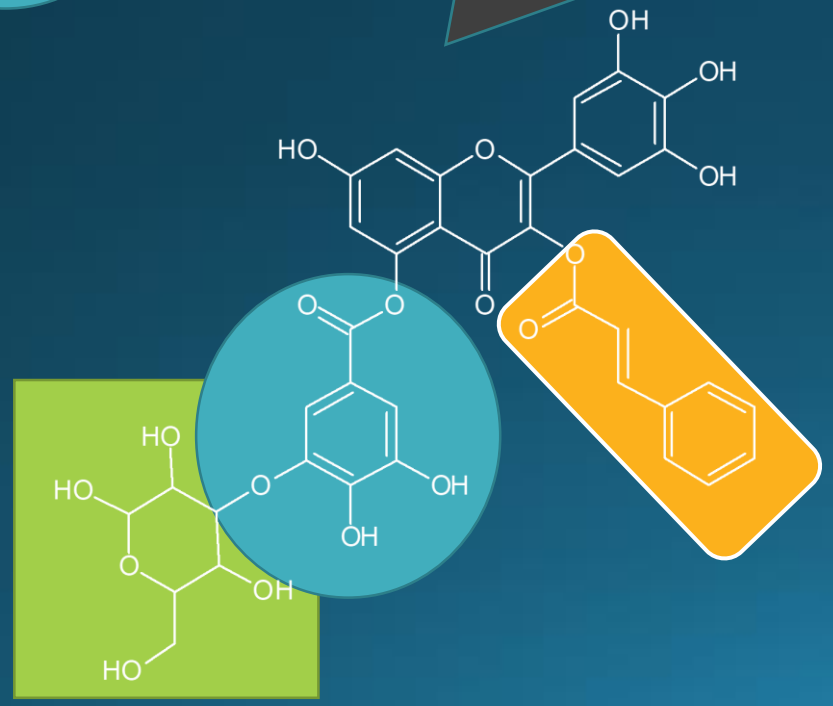


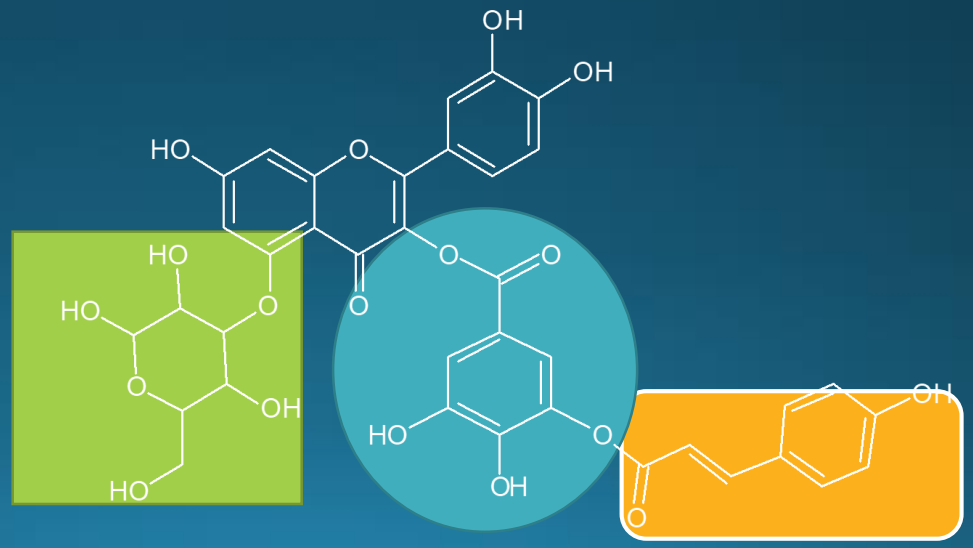


Combinatorial Chemistry

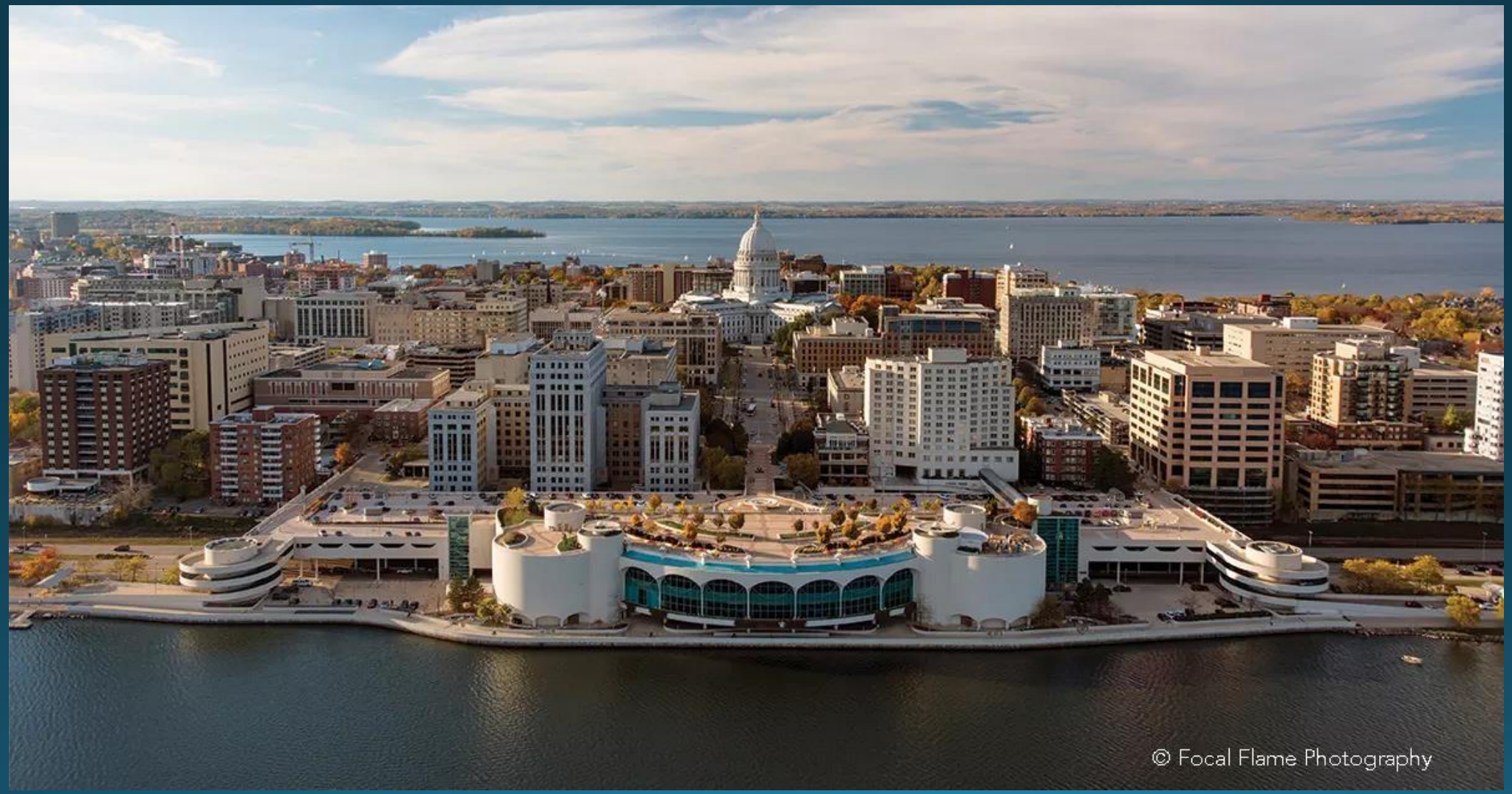






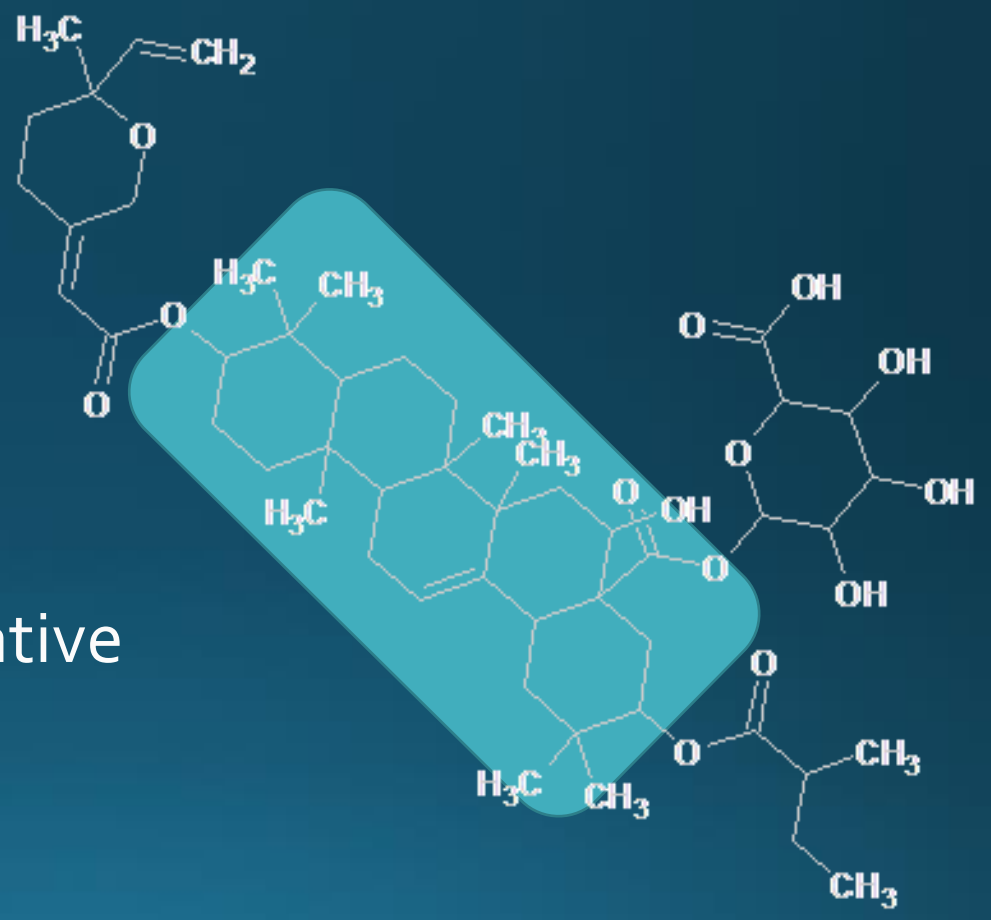


OSG User School 2017



Theoretical Saponins

- 7 Saponin Cores
- 21 Substituents
- Fully enumerated library
  - **4.2 million compounds…**
  - Reduced to 80,000
    - Fingerprints are not representative
  - Scaling back up



Acknowledgements

- Coley/Kursar Lab
  - Thomas Kursar
  - Lissy Coley
  - Dale Forrister
  - Gordon Younkin
- University of Utah CHPC
  - Wim Cardoen
  - Anita Orendt
  - Martin Čuma
- Open Science Grid
  - Bala Desinghu
  - Tim Cartwright
  - User School Personnel

Open Science Grid

NSF

CFM-ID

PEKING UNIVERSITY

Department of Biology
Bioscience Undergraduate Research Program
Center for High Performance Computing