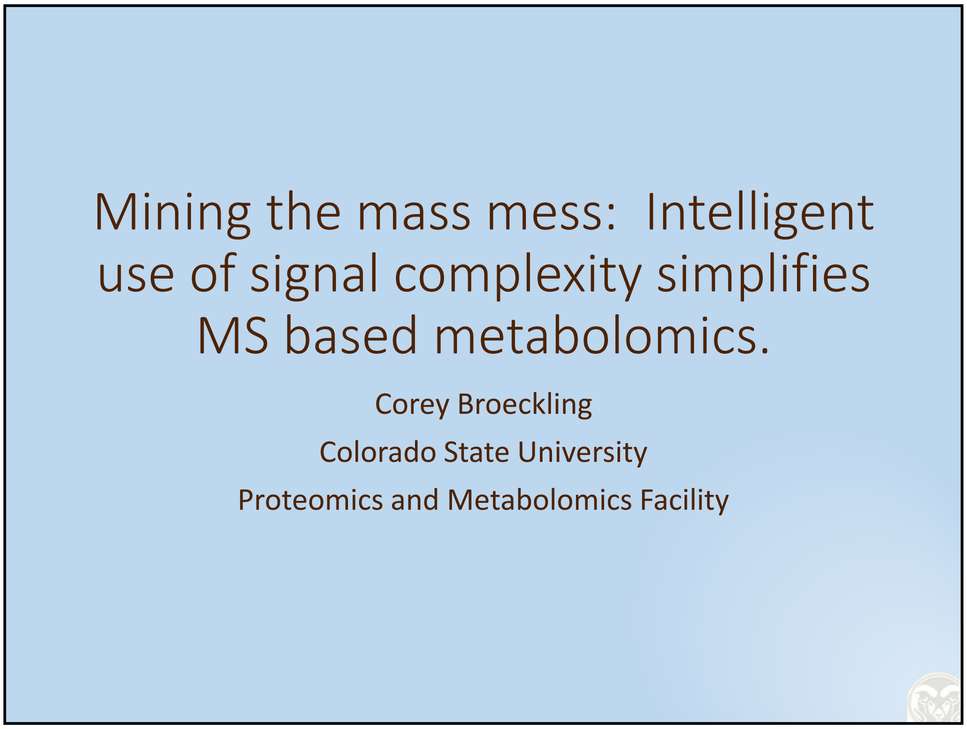
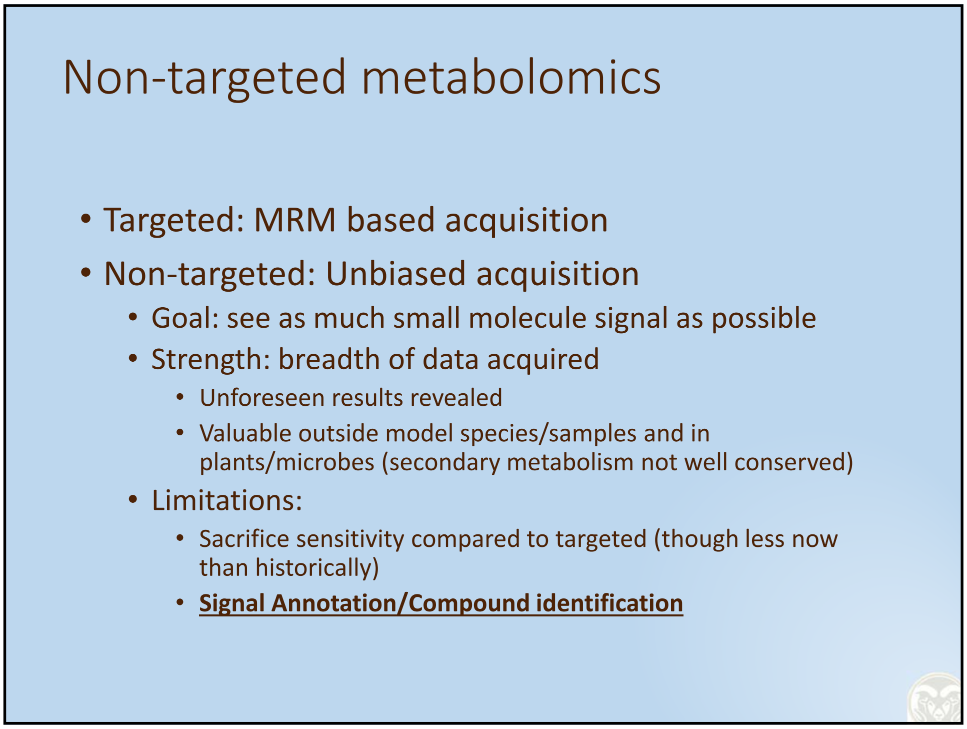


Mining the mass mess: Intelligent use of signal complexity simplifies MS based metabolomics.

Corey Broeckling

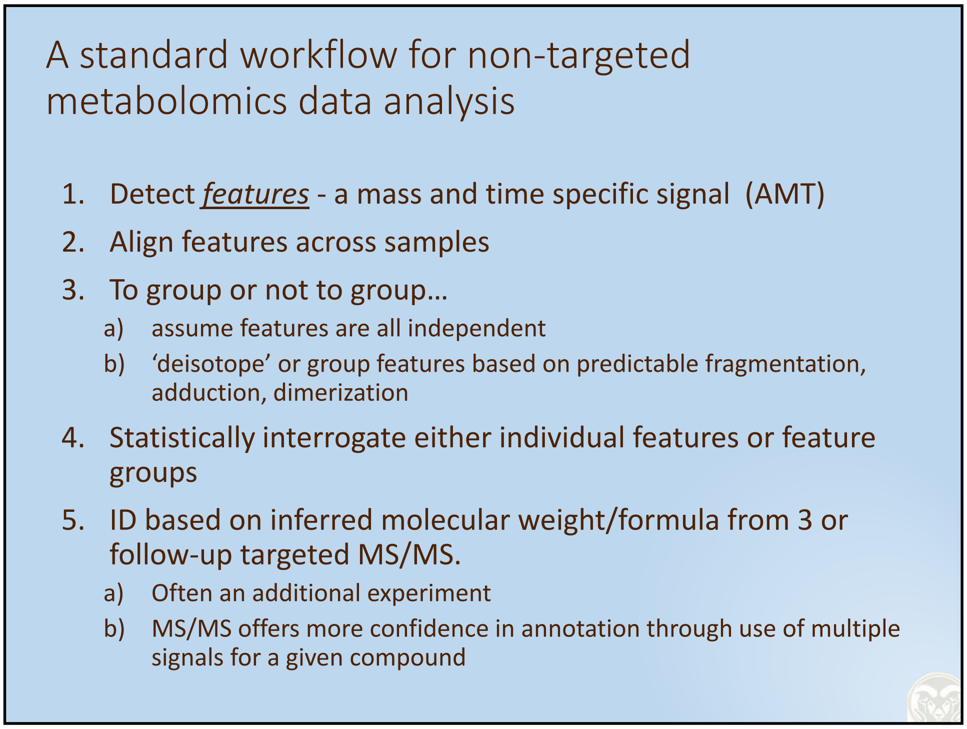
Colorado State University

Proteomics and Metabolomics Facility



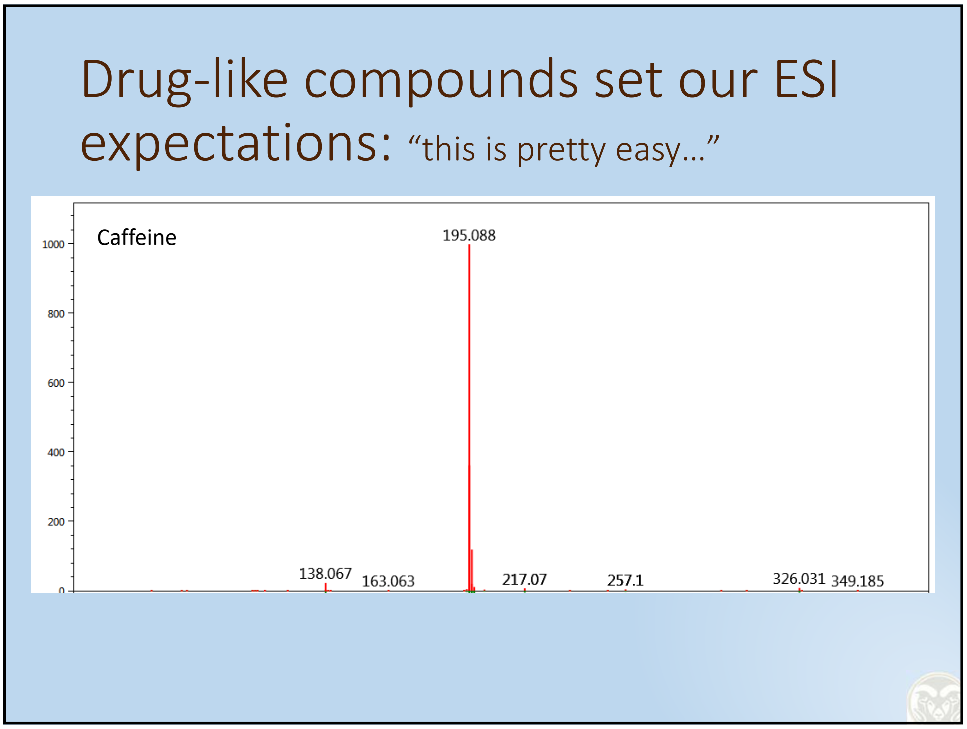
Non-targeted metabolomics

- Targeted: MRM based acquisition
- Non-targeted: Unbiased acquisition
  - Goal: see as much small molecule signal as possible
  - Strength: breadth of data acquired
    - Unforeseen results revealed
    - Valuable outside model species/samples and in plants/microbes (secondary metabolism not well conserved)
  - Limitations:
    - Sacrifice sensitivity compared to targeted (though less now than historically)
    - **Signal Annotation/Compound identification**



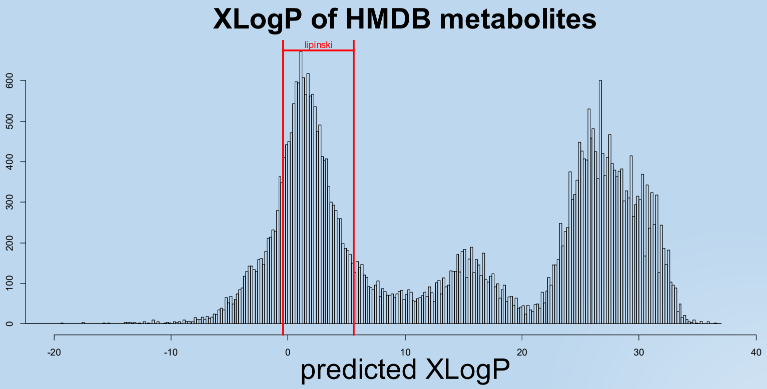
A standard workflow for non-targeted metabolomics data analysis

1. Detect *features* - a mass and time specific signal (AMT)
2. Align features across samples
3. To group or not to group…
   a) assume features are all independent
   b) 'deisotope' or group features based on predictable fragmentation, adduction, dimerization
4. Statistically interrogate either individual features or feature groups
5. ID based on inferred molecular weight/formula from 3 or follow-up targeted MS/MS.
   a) Often an additional experiment
   b) MS/MS offers more confidence in annotation through use of multiple signals for a given compound



Drug-like compounds set our ESI expectations: "this is pretty easy…"

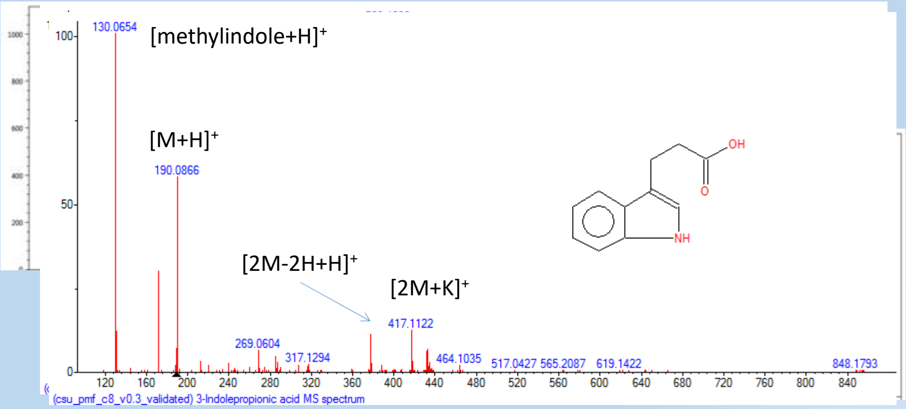
Most biological metabolites are not drug-like



- 73% of HMDB compounds have 1+ Lipinski failure(s)
- Diverse structure leads to diverse behavior

Diverse structures -> Diverse behavior:

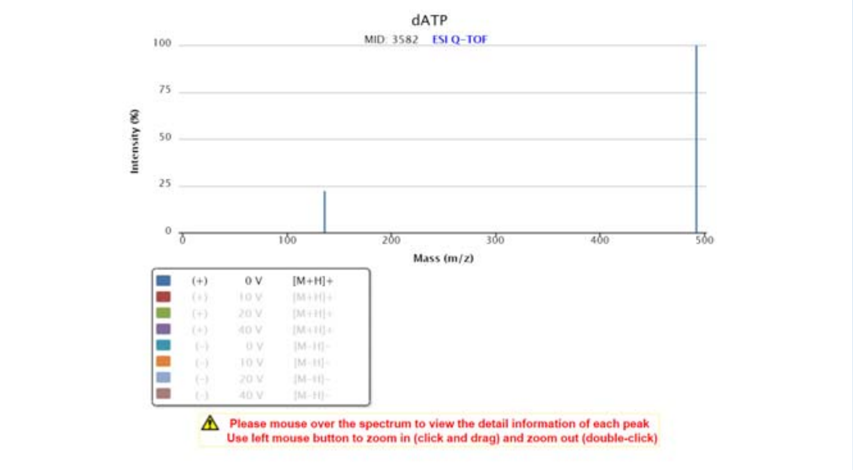
~1200 authentic standards run under real acquisition conditions – LC-TOF, positive ionization mode



(csu_pmf_c8_v0.3_validated) 3-Indolepropionic acid MS spectrum

- Some signals are easily predicted, others less so
- High risk of mis-intpretation!

In-source fragmentation happens

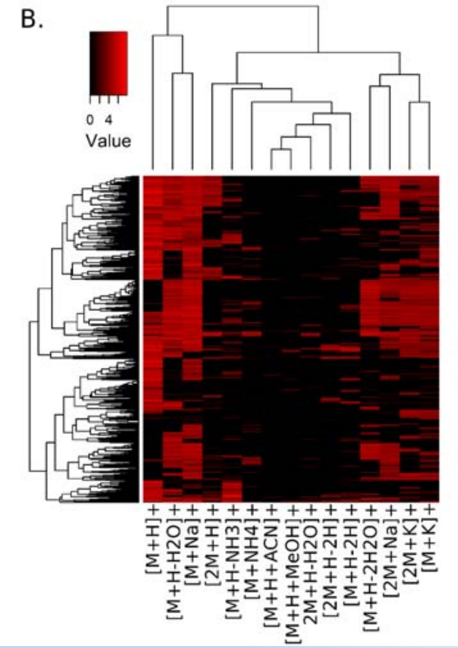


🟦	(+)	0 V	[M+H]+
🟥	(+)	10 V	[M+H]+
🟩	(+)	20 V	[M+H]+
🟪	(+)	40 V	[M+H]+
🟦	(-)	0 V	[M-H]-
🟧	(-)	10 V	[M-H]-
🟦	(-)	20 V	[M-H]-
🟫	(-)	40 V	[M-H]-

⚠ Please mouse over the spectrum to view the detail information of each peak
Use left mouse button to zoom in (click and drag) and zoom out (double-click)

Adduction can be complex

Rows = cmpds
Columns = adducts



Standard workflow for metabolomics data analysis

1. Detect features - a mass and time specific signal (AMT)
2. Align features across samples
3. To group or not to group…
   a) assume features are all independent
   b) 'deisotope' or group features based on predictable fragmentation, adduction, dimerization
4. Statistically interrogate either individual features or feature groups
5. ID based on follow-up MSMS or inferred molecular weight from 3.

Logical flaws in 3

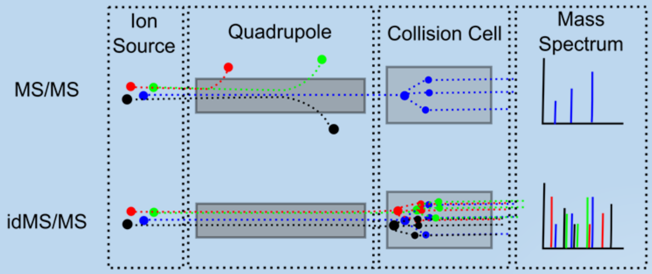
a) Features are not independent

b) Spectra are often unpredictable

- Implications of complex spectra:
  - Overestimation of sample complexity
  - Reduced spectral quality (and confidence) for known compounds
  - Wasted ID effort for redundant (identified) signals
- *SPECTRA ARE INFORMATIVE: diagnostic and interpretable!*
- *Feature grouping tools: AMDIS(NIST), MSClust(PRI), QUICS(Metabolon), Parafac2(University of Copenhagen), CAMERA (IPB-Halle, Germany)…*

Data Independent (MSe, MSall,…)

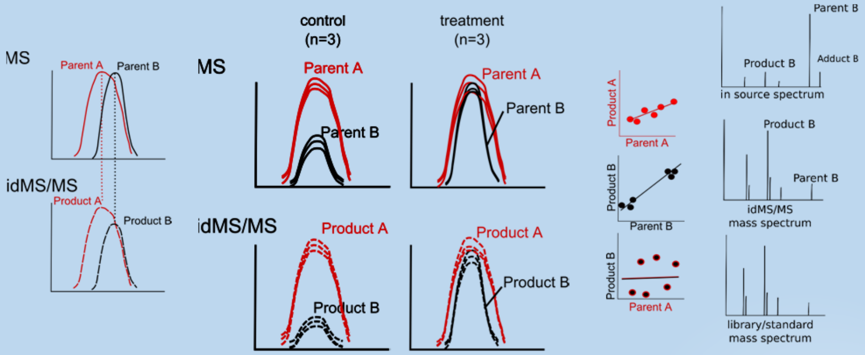
- MS^E: CID fragmentation without precursor isolation



- Concurrent acquisition, high and low collision energy
  - MS and MS/MS for all signals in single LC-MS injection
  - Issue: assigning precursor/product relationships

idMS/MS and in-source complexity:

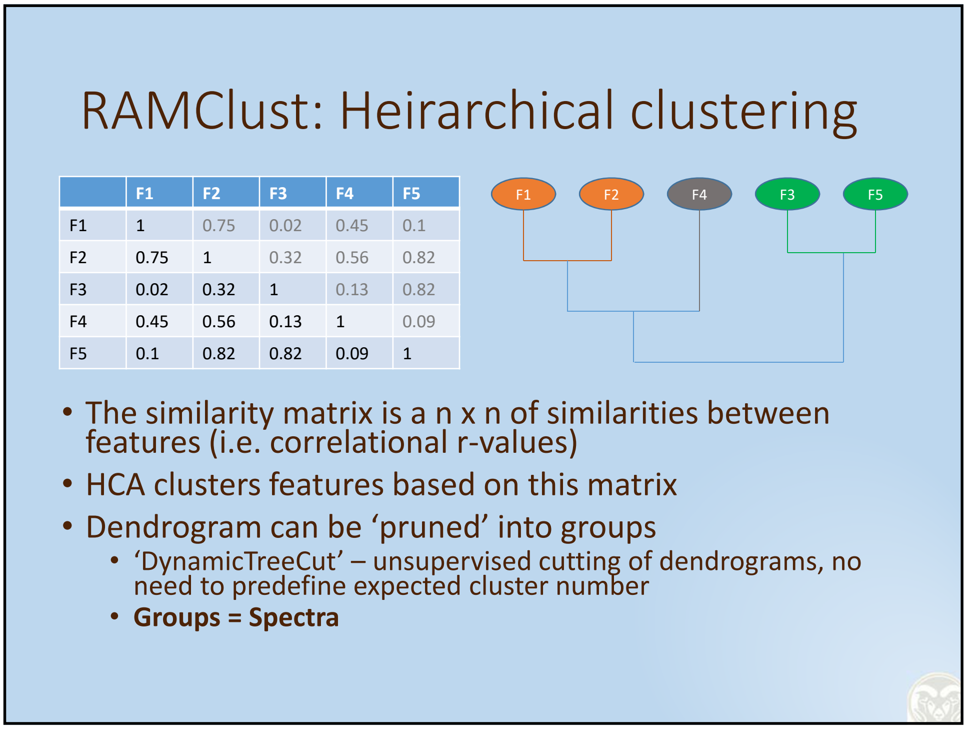
two problems, one solution



- Fragmentation/Adduction predictability unnecessary
- Integration reproducibility important
- Two parameters:
  - Retention time similarity
  - Correlational similarity

RAMClustR: custom similarity matrix

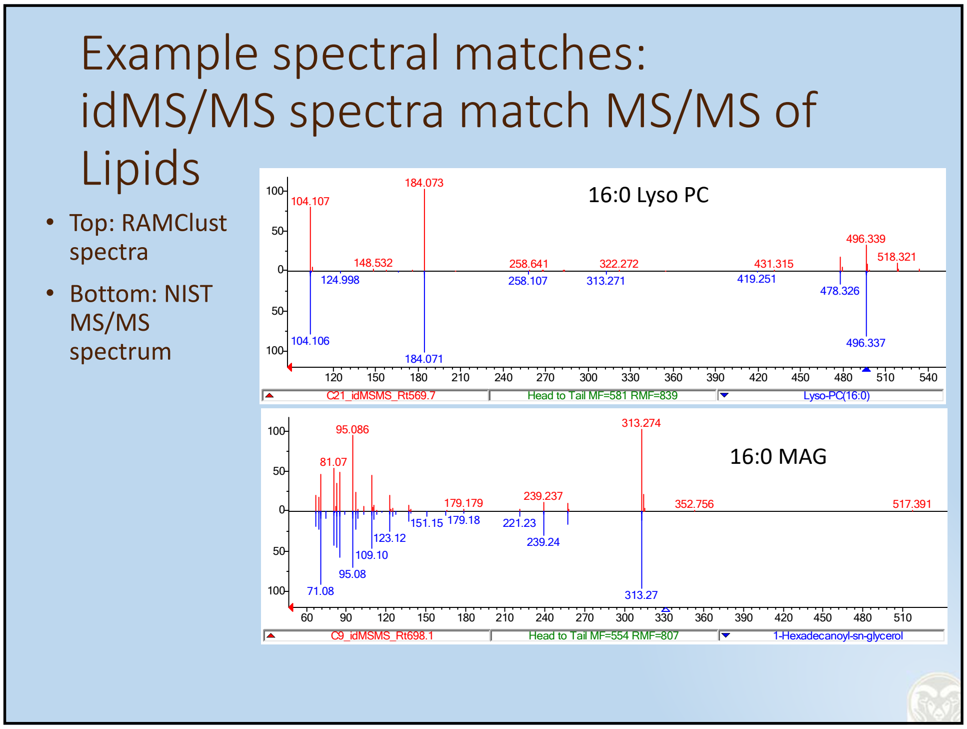
- Similarity between two features is the product of two gaussian functions (σ is tunable in each)
  - Correlation (quantitative similarity [r], MS vs MS, MS vs MS/MS, MS/MS vs MS/MS)
  - Retention time (temporal coelution)
  - No cutoffs!
- If either correlation or retention time is *dissimilar*, total similarity approaches zero
- Similarity calculated for all pairs of features



RAMClust: Heirarchical clustering

	F1	F2	F3	F4	F5
F1	1	0.75	0.02	0.45	0.1
F2	0.75	1	0.32	0.56	0.82
F3	0.02	0.32	1	0.13	0.82
F4	0.45	0.56	0.13	1	0.09
F5	0.1	0.82	0.82	0.09	1

- The similarity matrix is a n x n of similarities between features (i.e. correlational r-values)
- HCA clusters features based on this matrix
- Dendrogram can be ‘pruned’ into groups
  - ‘DynamicTreeCut’ – unsupervised cutting of dendrograms, no need to predefine expected cluster number
  - **Groups = Spectra**



Example spectral matches: idMS/MS spectra match MS/MS of Lipids

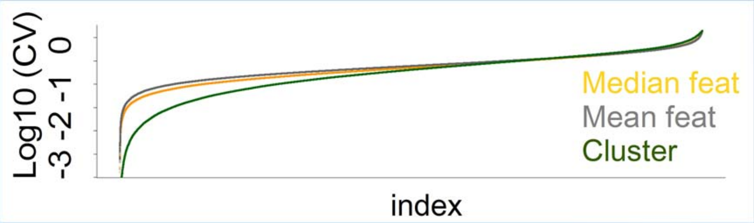
- Top: RAMClust spectra
- Bottom: NIST MS/MS spectrum

RAMClust overview:

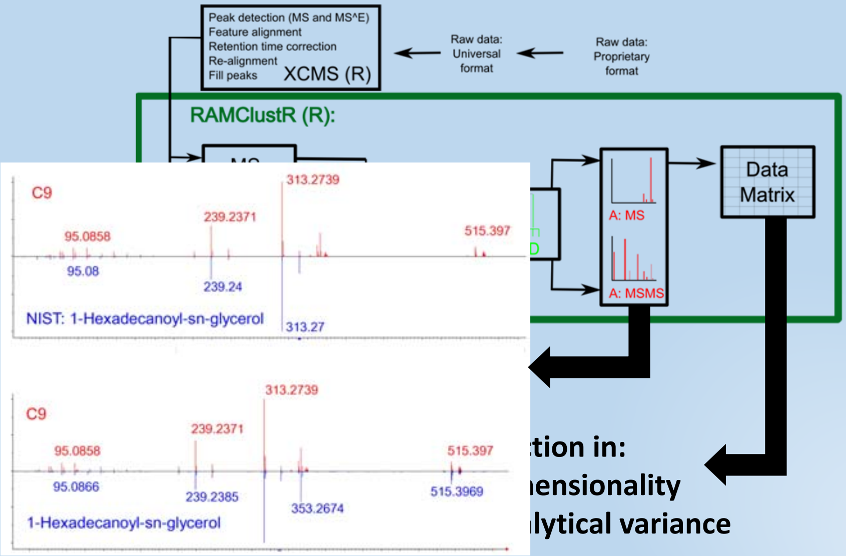
- Only input: dataset(s)
- Independent of predictability in:
  - Adduction
  - Fragmentation
  - Isotope pattern
- Output to .msp format
  - *Both MS and idMS/MS spectra*
  - Viewing and searching using MSSearch (NIST)
- Fast
  - Dataset(s) of 17,000 features to msp spectral library < 200 seconds
  - Easier downstream: 17,000 features ~ 2700 clusters
- Spectral searching
  - More reliable than MW – multiple signals!
  - No assumptions regarding MW of compound
  - Spectral searching offers a shallow learning curve compared to spectral interpretation
- Spectra more interpretable than features
- Dependent on reliable feature detection and integration

Bonus: reduced variance when using spectra

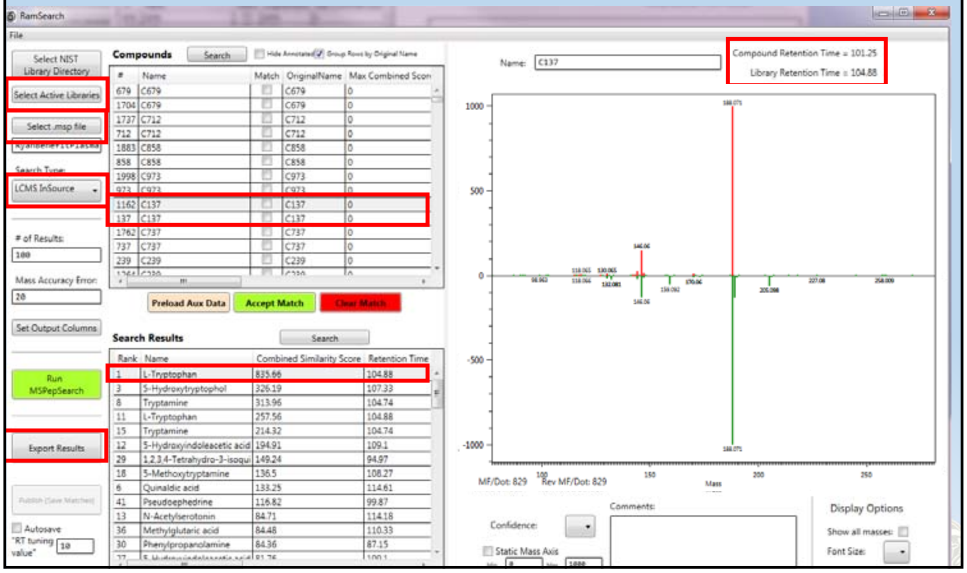
- Redundant measures of metabolite abundance results in a lower CV
- Lower analytical CV -> better sensitivity to biological changes



RAMClustR: data driven feature clustering



RAMSearch: metabolomics-centered spectral search GUI



Search Type: LCMS InSource

# of Results: 100

Mass Accuracy Error: 20

□ Autosave

"RT tuning value" 10

#	Name	Match	OriginalName	Max Combined Score
679	C679	□	C679	0
1704	C679	□	C679	0
1737	C712	□	C712	0
712	C712	□	C712	0
1883	C858	□	C858	0
858	C858	□	C858	0
1998	C973	□	C973	0
973	C973	□	C973	0
1162	C137	□	C137	0
137	C137	□	C137	0
1762	C737	□	C737	0
737	C737	□	C737	0
239	C239	□	C239	0
1264	C239	□	C239	0

Rank	Name	Combined Similarity Score	Retention Time
1	L-Tryptophan	835.66	104.88
3	5-Hydroxytryptophol	326.19	107.33
8	Tryptamine	313.96	104.74
11	L-Tryptophan	257.56	104.88
15	Tryptamine	214.32	104.74
12	5-Hydroxyindoleacetic acid	194.91	109.1
29	1,2,3,4-Tetrahydro-3-isoqui	149.24	94.97
18	5-Methoxytryptamine	136.5	108.27
6	Quinaldic acid	133.25	114.61
41	Pseudoephedrine	116.82	99.87
13	N-Acetylserotonin	84.71	114.18
36	Methylglutaric acid	84.48	110.33
30	Phenylpropanolamine	84.36	87.15
27	5-Hydroxyindoleacetic acid	81.76	100.1

Name: C137

Compound Retention Time = 101.25

Library Retention Time = 104.88

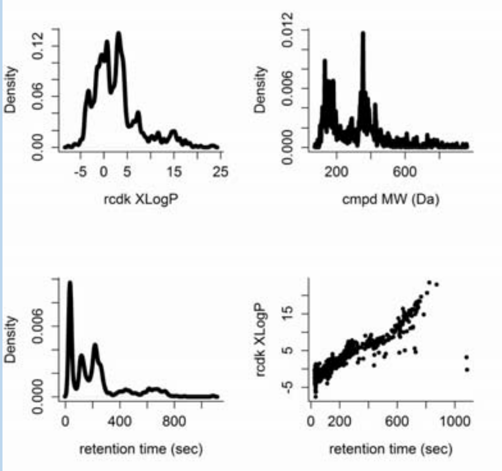
MF/Dot: 829 Rev MF/Dot: 829

□ Static Mass Axis Min 0 Max 1000

Compound # >>> Spectra

- METLIN: 961,829 compounds
  - ~ 14,000 with authentic spectra (image format)
- LipidMaps: ~40,000 compounds
  - ~ 500 with spectra
- PubChem: 93,553,257 compounds
  - NIST
    - GC-MS EI spectra 267,376 compounds
    - LC-MS/MS spectra 14,351 compounds
- **Authentic standard spectral libraries will be incomplete for the foreseeable future**
  - **Can predicted analytical behavior help?**

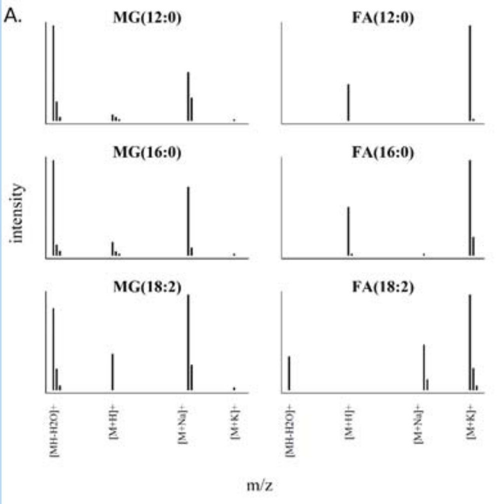
CSU PMF Spectral and Retention Time Library



- ~ 900 authentic standards
- UPLC c8 reverse phase MeOH gradient
  - Phenyl Hexyl ACN
  - HILIC
- MS and MSE data for each compound
- Retention time for each compound

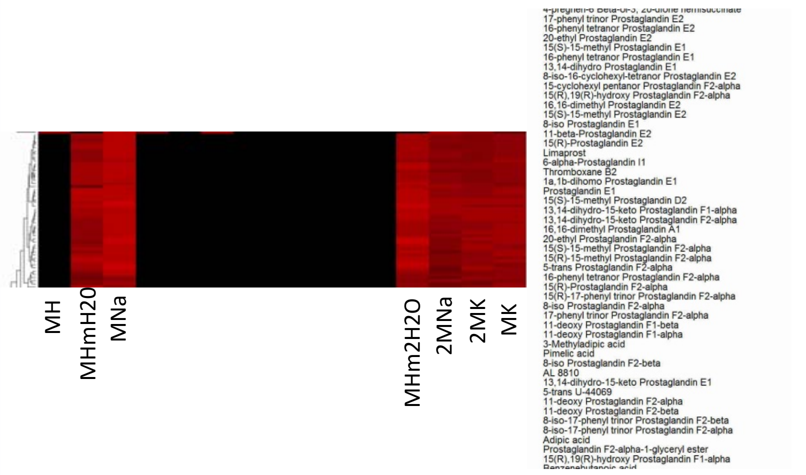
Adduction is not random

patterns contain information diagnostic of structure



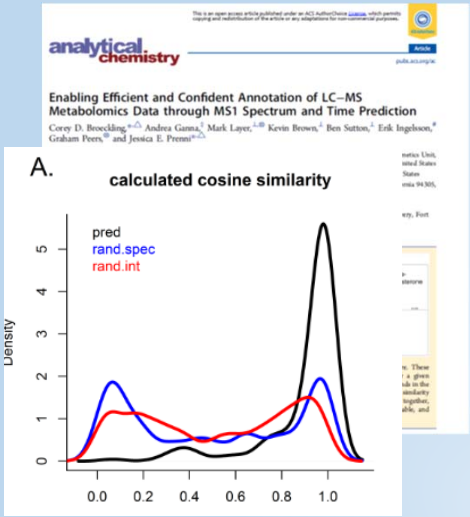
Adduction is not random

patterns contain information diagnostic of structure



In-source patterns are predictable

- Structure database converted to retention time and MS level spectral library
- No MS/MS necessary (though will often be beneficial)
- Utilize MS signal 'redundancy' for efficient and confident annotation!
- Approach doesn't scale well



This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

analytical chemistry

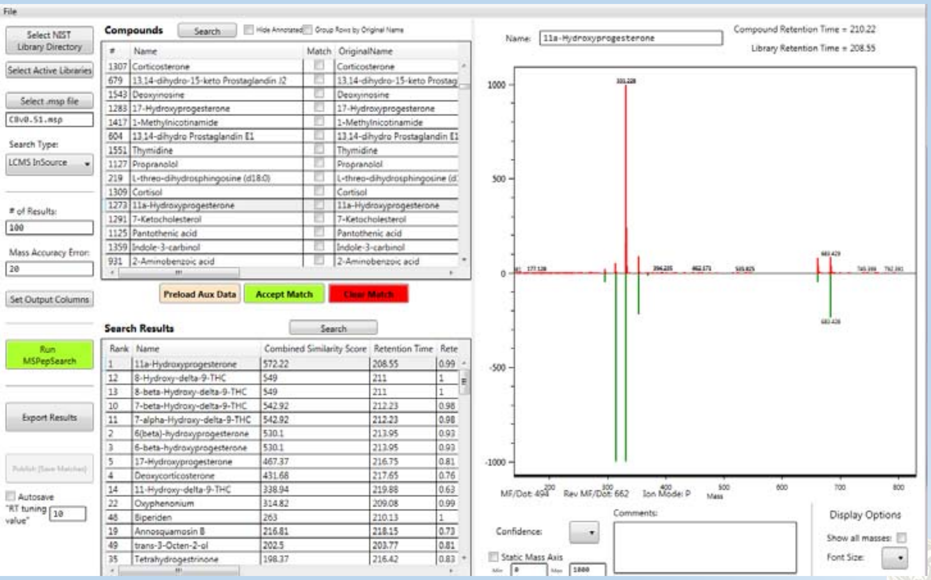
Article

pubs.acs.org/ac

Enabling Efficient and Confident Annotation of LC–MS Metabolomics Data through MS1 Spectrum and Time Prediction

Corey D. Broeckling, Andrea Ganna, Mark Layer, Kevin Brown, Ben Sutton, Erik Ingelsson, Graham Peers, and Jessica E. Prenni

In-source spectra are searchable



C8v0.51.msp

Search Type: LCMS InSource

# of Results: 100

Mass Accuracy Error: 20

☐ Autosave

"RT tuning value" 10

☐ Hide Annotated ☐ Group Rows by Original Name

#	Name	Match	OriginalName
1307	Corticosterone	☐	Corticosterone
679	13,14-dihydro-15-keto Prostaglandin J2	☐	13,14-dihydro-15-keto Prostag
1543	Deoxyinosine	☐	Deoxyinosine
1283	17-Hydroxyprogesterone	☐	17-Hydroxyprogesterone
1417	1-Methylnicotinamide	☐	1-Methylnicotinamide
604	13,14-dihydro Prostaglandin E1	☐	13,14-dihydro Prostaglandin E1
1551	Thymidine	☐	Thymidine
1127	Propranolol	☐	Propranolol
219	L-threo-dihydrosphingosine (d18:0)	☐	L-threo-dihydrosphingosine (d
1309	Cortisol	☐	Cortisol
1273	11a-Hydroxyprogesterone	☐	11a-Hydroxyprogesterone
1291	7-Ketocholesterol	☐	7-Ketocholesterol
1125	Pantothenic acid	☐	Pantothenic acid
1359	Indole-3-carbinol	☐	Indole-3-carbinol
931	2-Aminobenzoic acid	☐	2-Aminobenzoic acid

Rank	Name	Combined Similarity Score	Retention Time	Rete
1	11a-Hydroxyprogesterone	572.22	208.55	0.99
12	8-Hydroxy-delta-9-THC	549	211	1
13	8-beta-Hydroxy-delta-9-THC	549	211	1
10	7-beta-Hydroxy-delta-9-THC	542.92	212.23	0.98
11	7-alpha-Hydroxy-delta-9-THC	542.92	212.23	0.98
2	6(beta)-hydroxyprogesterone	530.1	213.95	0.93
3	6-beta-hydroxyprogesterone	530.1	213.95	0.93
5	17-Hydroxyprogesterone	467.37	216.75	0.81
4	Deoxycorticosterone	431.68	217.65	0.76
14	11-Hydroxy-delta-9-THC	338.94	219.88	0.63
22	Oxyphenonium	314.82	209.08	0.99
48	Biperiden	263	210.13	1
19	Annosquamosin B	216.81	218.15	0.73
49	trans-3-Octen-2-ol	202.5	203.77	0.81
35	Tetrahydrogestrinone	198.37	216.42	0.83

Name: 11a-Hydroxyprogesterone

Compound Retention Time = 210.22

Library Retention Time = 208.55

MF/Dot: 494 Rev MF/Dot: 662 Ion Mode: P

Comments:

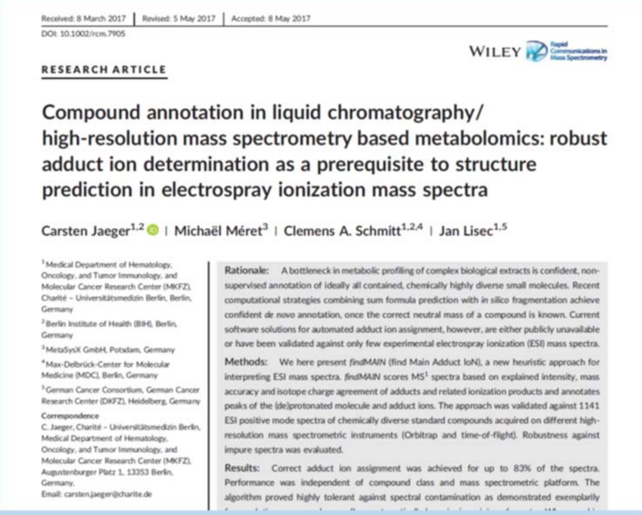
Confidence:

☐ Static Mass Axis Min 0 Max 1000

Show all masses: ☐

Font Size:

In-source spectra are interpretable



Received: 8 March 2017 | Revised: 5 May 2017 | Accepted: 8 May 2017
DOI: 10.1002/rcm.7905

WILEY Rapid Communications in Mass Spectrometry

RESEARCH ARTICLE

Compound annotation in liquid chromatography/high-resolution mass spectrometry based metabolomics: robust adduct ion determination as a prerequisite to structure prediction in electrospray ionization mass spectra

Carsten Jaeger[1,2] | Michaël Méret[3] | Clemens A. Schmitt[1,2,4] | Jan Lisec[1,5]

[1] Medical Department of Hematology, Oncology, and Tumor Immunology, and Molecular Cancer Research Center (MKFZ), Charité – Universitätsmedizin Berlin, Berlin, Germany

[2] Berlin Institute of Health (BIH), Berlin, Germany

[3] MetaSysX GmbH, Potsdam, Germany

[4] Max-Delbrück-Center for Molecular Medicine (MDC), Berlin, Germany

[5] German Cancer Consortium, German Cancer Research Center (DKFZ), Heidelberg, Germany

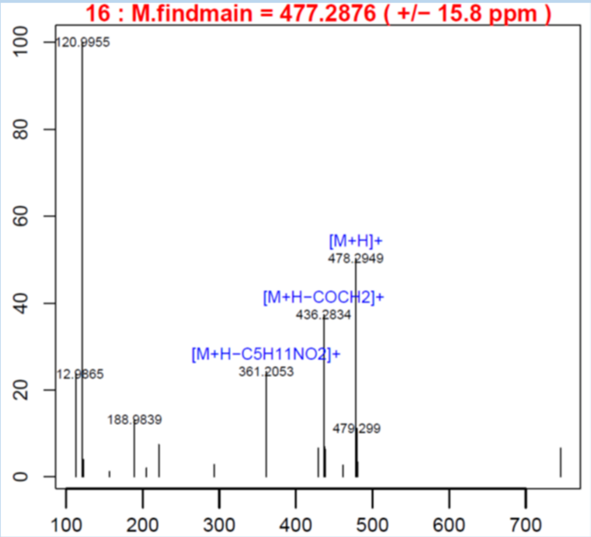
Correspondence
C. Jaeger, Charité – Universitätsmedizin Berlin, Medical Department of Hematology, Oncology, and Tumor Immunology, and Molecular Cancer Research Center (MKFZ), Augustenburger Platz 1, 13353 Berlin, Germany.
Email: carsten.jaeger@charite.de

Rationale: A bottleneck in metabolic profiling of complex biological extracts is confident, non-supervised annotation of ideally all contained, chemically highly diverse small molecules. Recent computational strategies combining sum formula prediction with *in silico* fragmentation achieve confident *de novo* annotation, once the correct neutral mass of a compound is known. Current software solutions for automated adduct ion assignment, however, are either publicly unavailable or have been validated against only few experimental electrospray ionization (ESI) mass spectra.

Methods: We here present *findMAIN* (find Main Adduct IoN), a new heuristic approach for interpreting ESI mass spectra. *findMAIN* scores MS^1 spectra based on explained intensity, mass accuracy and isotope charge agreement of adducts and related ionization products and annotates peaks of the (de)protonated molecule and adduct ions. The approach was validated against 1141 ESI positive mode spectra of chemically diverse standard compounds acquired on different high-resolution mass spectrometric instruments (Orbitrap and time-of-flight). Robustness against impure spectra was evaluated.

Results: Correct adduct ion assignment was achieved for up to 83% of the spectra. Performance was independent of compound class and mass spectrometric platform. The algorithm proved highly tolerant against spectral contamination as demonstrated exemplarily

In-source spectra are interpretable

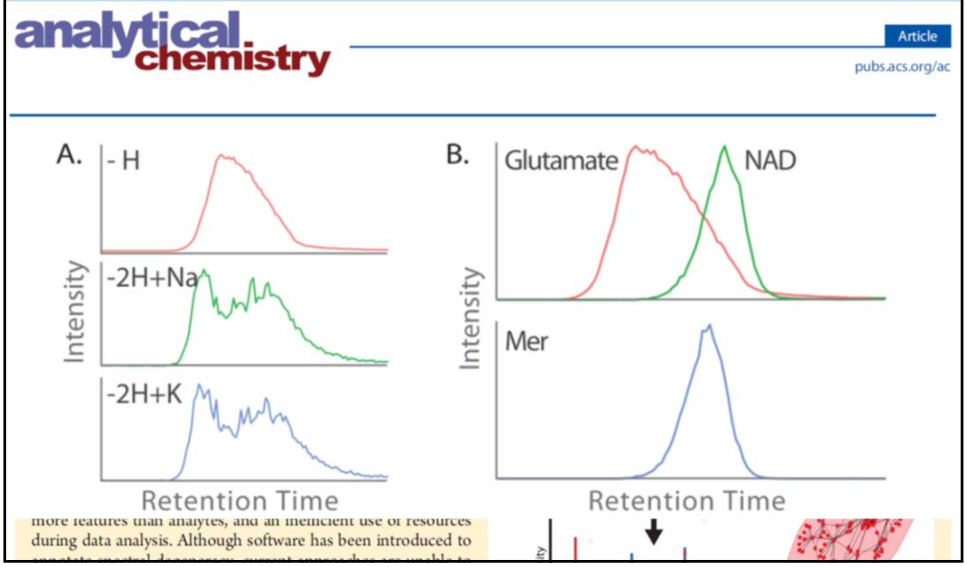


16 : M.findmain = 477.2876 (+/− 15.8 ppm)

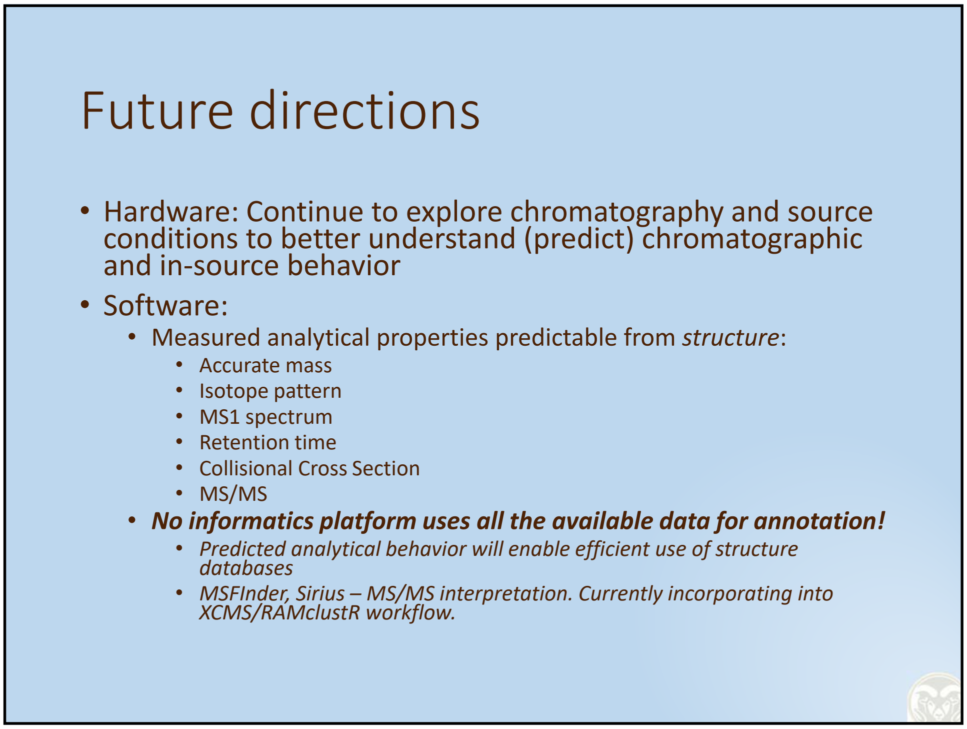
Summary: In-Source Phenomenon

- Biological metabolites often generate complex MS spectra
  - Features can be grouped without chemical assumptions
    - RAMClustR
  - In-source spectral complexity can be useful
    - In-source spectral matching using
      - NIST Search tools
      - RAMSearch
    - 1-SToP approach
      - Predicted in-source spectrum and retention time
      - Theoretical MS & RT signals from chemical structures
        - HMDB, LipidMaps….
      - RAMSearch
    - Interpretation of in-source delta mass to obtain more confident molecular weight – interpretMSSpectrum (called from RAMClustR)

Heterodimers !???!?! Ackkk!?!!

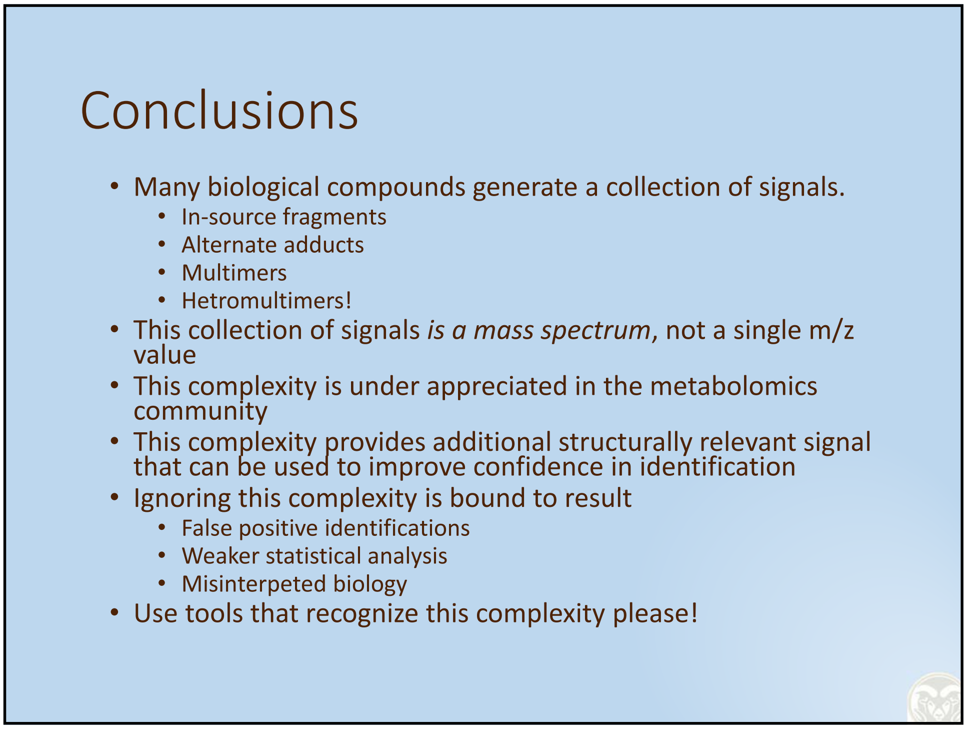


more features than analytes, and an inefficient use of resources during data analysis. Although software has been introduced to



Future directions

- Hardware: Continue to explore chromatography and source conditions to better understand (predict) chromatographic and in-source behavior
- Software:
  - Measured analytical properties predictable from *structure*:
    - Accurate mass
    - Isotope pattern
    - MS1 spectrum
    - Retention time
    - Collisional Cross Section
    - MS/MS
  - ***No informatics platform uses all the available data for annotation!***
    - *Predicted analytical behavior will enable efficient use of structure databases*
    - *MSFInder, Sirius – MS/MS interpretation. Currently incorporating into XCMS/RAMclustR workflow.*

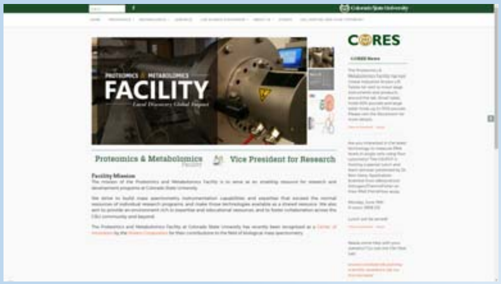


Conclusions

- Many biological compounds generate a collection of signals.
  - In-source fragments
  - Alternate adducts
  - Multimers
  - Hetromultimers!
- This collection of signals *is a mass spectrum*, not a single m/z value
- This complexity is under appreciated in the metabolomics community
- This complexity provides additional structurally relevant signal that can be used to improve confidence in identification
- Ignoring this complexity is bound to result
  - False positive identifications
  - Weaker statistical analysis
  - Misinterpeted biology
- Use tools that recognize this complexity please!

Acknowledgements

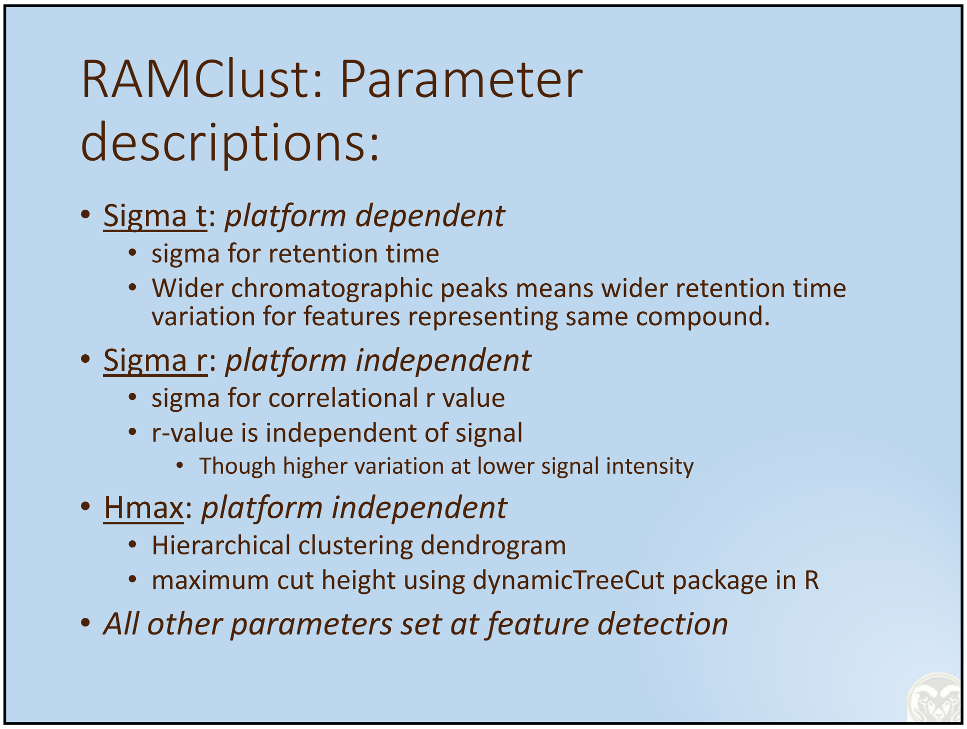
- CSU
  - Jessica Prenni (PMF)
  - PMF lab
  - Kevin Brown, Ben Sutton, Graham Peers, and Mark Layer
    - RAMSearch
- Broad Institute
  - Andrea Ganna
- Stanford
  - Erik Ingellson


Colorado State University
PROTEOMICS & METABOLOMICS
FACILITY
Local Discovery Global Impact
CORES
CORES News
Proteomics & Metabolomics Facility
Vice President for Research
Facility Mission

RAMClust: custom similarity matrix

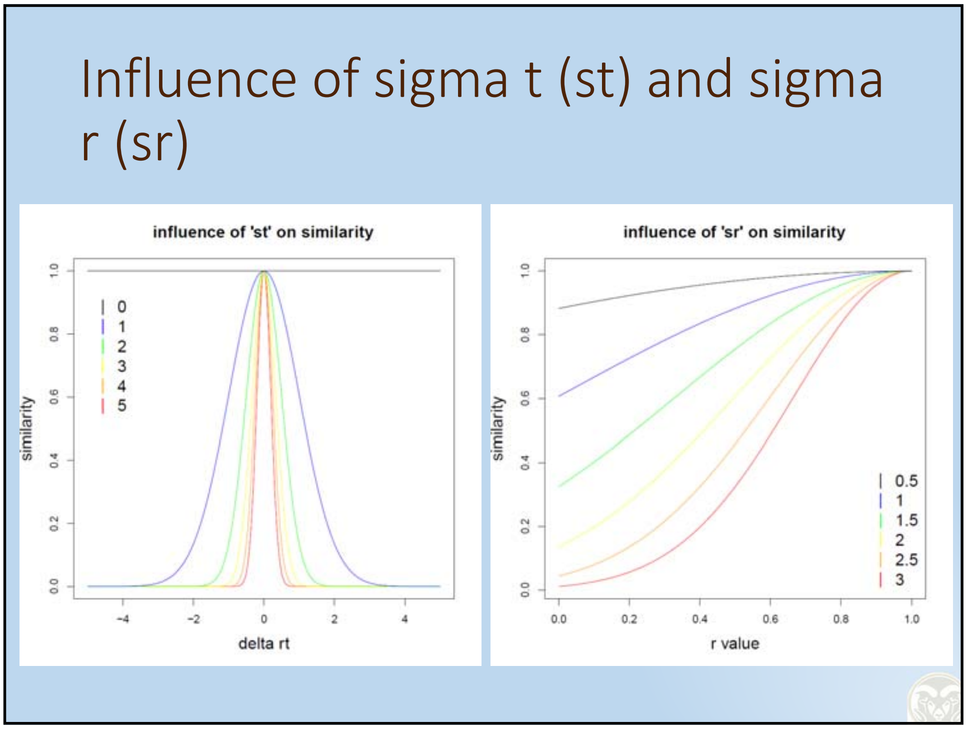
$$S_{ij} = \boxed{\frac{1}{\alpha}\begin{pmatrix} \alpha_1 e^{-\left(1-c_{ij}^{MS1/MS1}\right)^2/2\sigma_1^2} + \\ \alpha_2 e^{-\left(1-c_{ij}^{MS2/MS2}\right)^2/2\sigma_2^2} + \\ \alpha_{12} e^{-\left(1-c_{ij}^{MS1/MS2}\right)^2/2\sigma_{12}^2} \end{pmatrix}}\boxed{e^{-\left(t_i-t_j\right)^2/2\sigma_t^2}}$$

- Similarity between two features is the product of two gaussian functions (σ is tunable in each)
  - Correlation (quantitative similarity, MS-MS, MS-MS/MS, MS/MS-MS/MS)
  - Retention time (temporal coelution)
  - No cutoffs!
- If either cor or rt is *dissimilar*, total similarity approaches zero
- Use Data Dependent MS/MS precursor-product relationships to examine quality of clustering
  - Response: average spectral similarity for all feature-mapped DDA spectra which have similarity > 0.5 for ANY combination of parameters



RAMClust: Parameter descriptions:

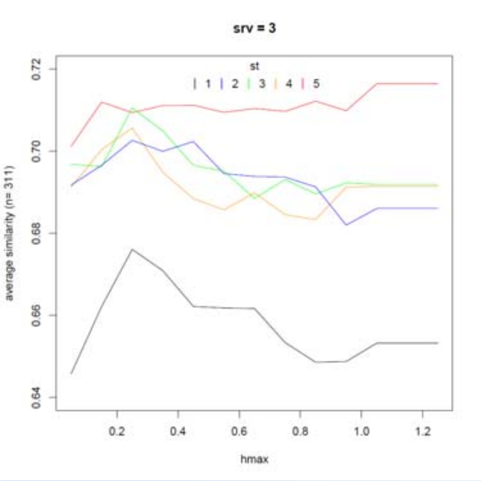
- Sigma t: *platform dependent*
  - sigma for retention time
  - Wider chromatographic peaks means wider retention time variation for features representing same compound.
- Sigma r: *platform independent*
  - sigma for correlational r value
  - r-value is independent of signal
    - Though higher variation at lower signal intensity
- Hmax: *platform independent*
  - Hierarchical clustering dendrogram
  - maximum cut height using dynamicTreeCut package in R
- *All other parameters set at feature detection*

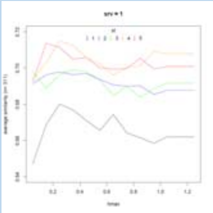


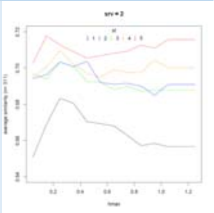
Influence of sigma t (st) and sigma r (sr)

DDA MS/MS spectra validate RAMClust relationships:

- Use known precursor-product relationships as the benchmark
- Any sim > 0.5
  - *n=311 DDA MS/MS*
- 390 combinations
- Y-axis: average spectral similarity (from 0-1)
- Stability to hmax improves at higher sigma_r and sigma_t

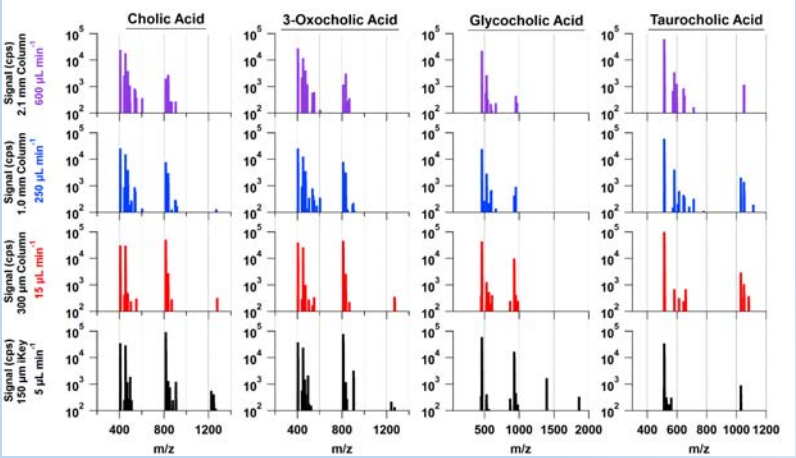


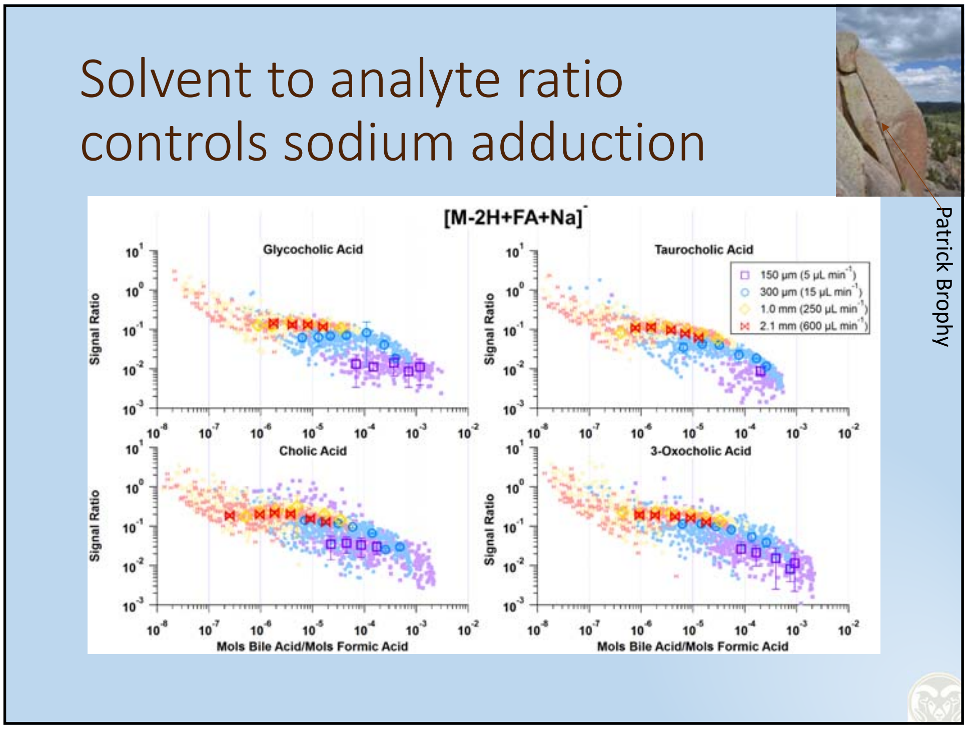




Can we control in-source complexity?

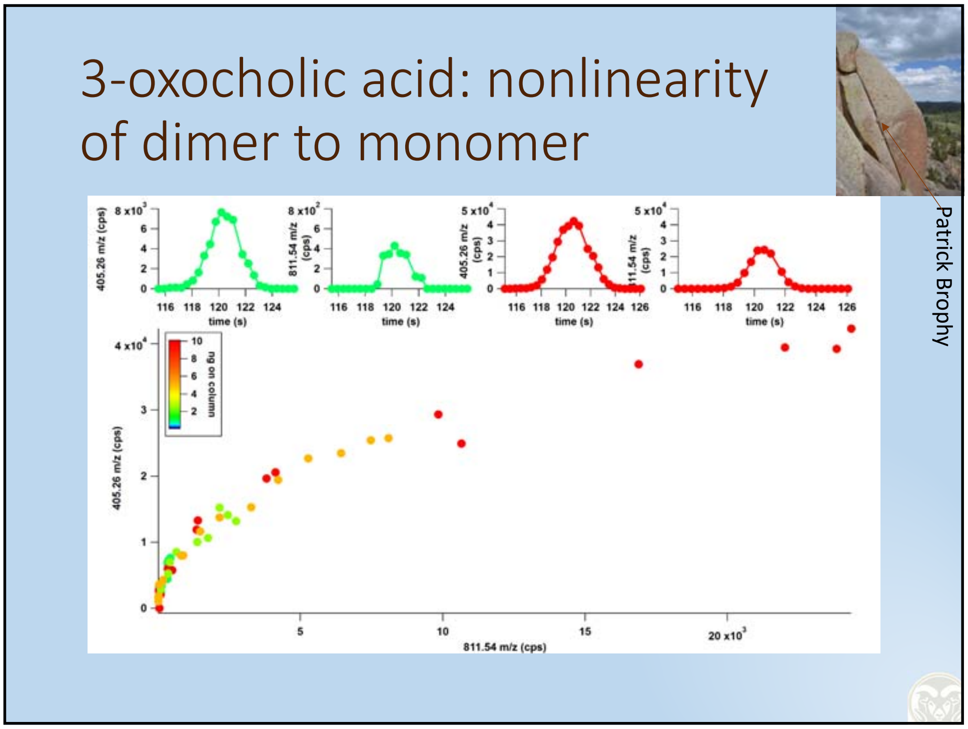
Patrick Brophy





Solvent to analyte ratio controls sodium adduction

Patrick Brophy

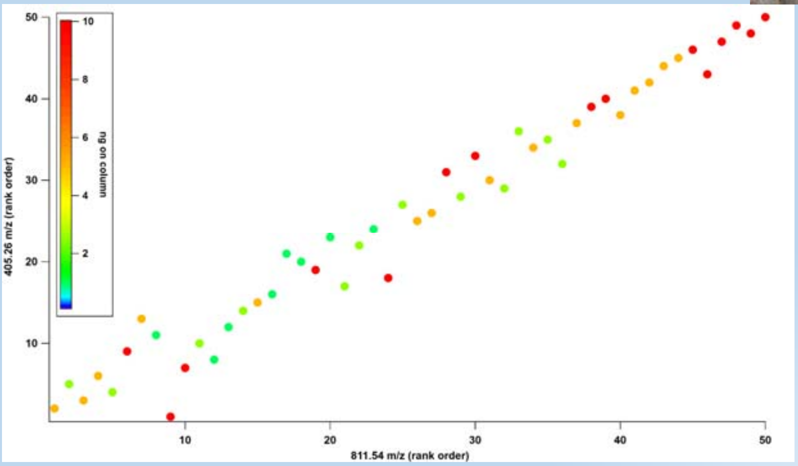


3-oxocholic acid: nonlinearity of dimer to monomer

Patrick Brophy

3-oxocholic acid: ranked intensity shows strong linear relationship

Patrick Brophy


405.26 m/z (rank order)
811.54 m/z (rank order)
ng on column