Supplementary Information

Systems-Level Annotation of a Metabolomics Data Set Reduces 25,000 Features to Fewer than 1,000 Unique Metabolites

Nathaniel G. Mahieu and Gary J. Patti*

Department of Chemistry, Washington University, St. Louis, Missouri 63130, United States

*Contact: gjpattij@wustl.edu, 314-935-3512

Table S1. Feature counts detected with various XCMS parameters.

Feature Count	method	ppm	peakwidth	prefilter	snthresh
56,368	“centWave”	6	c(2, 8)	c(4, 1000)	10
36,523	“centWave”	6	c(4, 8)	c(4, 1000)	20

Table S2. A breakdown of the analyte number observed after each annotation step.

	Groups with more than one feature		Singlets	
Stage	**All Features**	**Credentialed Features**	**All Features**	**Credentialed Features**
Blank Subtracted	0	0	12797	2462
Isotopes	3986	1066	5071	1326
Charge Carriers	3620	1137	4384	992
Neutral Losses	3640	1174	3678	790
Multimers	3400	1117	3381	712
Commons n>200	2809	1063	2472	495
Commons n>50	2149	864	1620	353
Background	1673	659	1288	233

Figure S1. An overview of the consensus data set. (A) The base peak chromatogram of a representative run. The number of features detected during each second is overlaid. (B) The number of features detected in each group before (pink) and after (green) Warpgroup. Inconsistencies are resolved by Warpgroup. (C) The within group CVs of peak areas is decreased by Warpgroup. (D) The within group CVs of peak width are decreased by Warpgroup. (E) Several representative features detected by the informatic workflow. The estimated baseline is plotted in red.

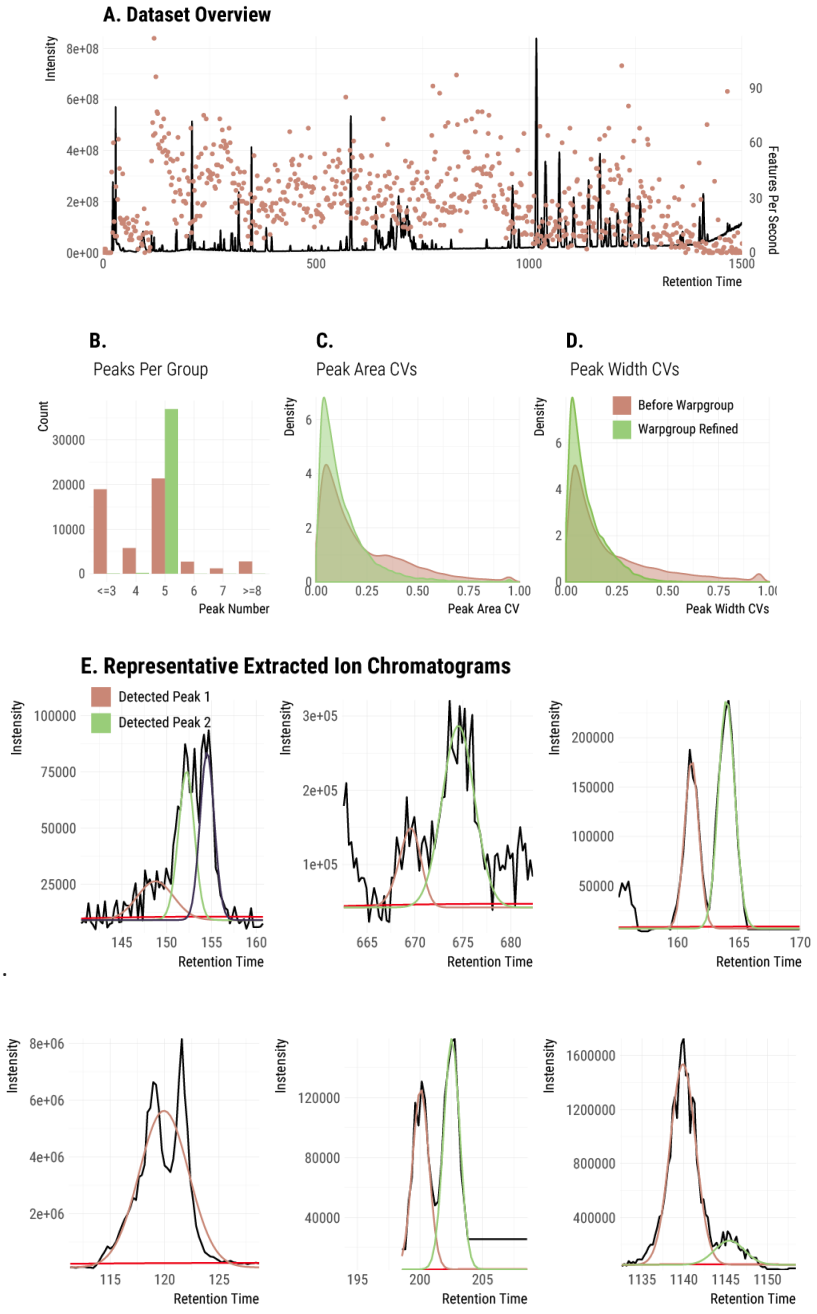
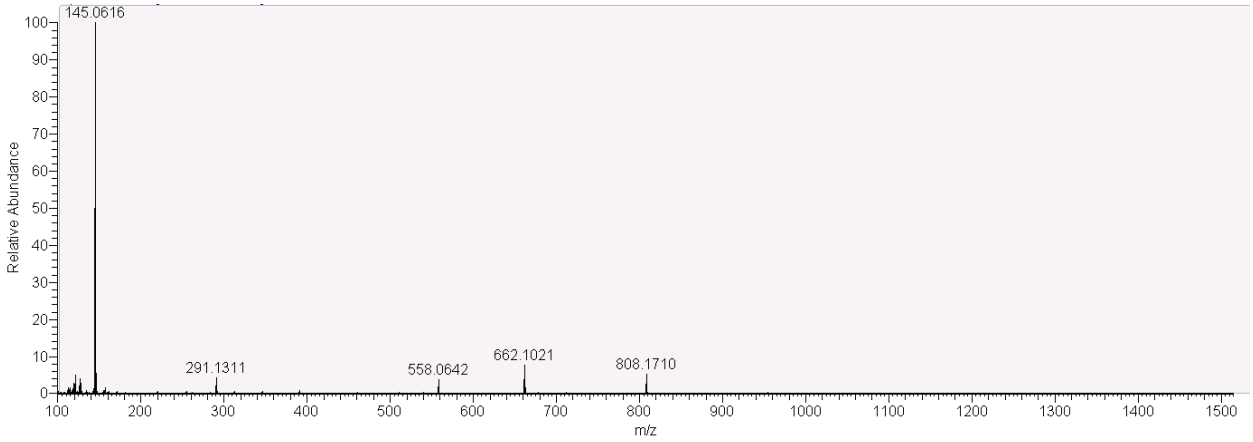
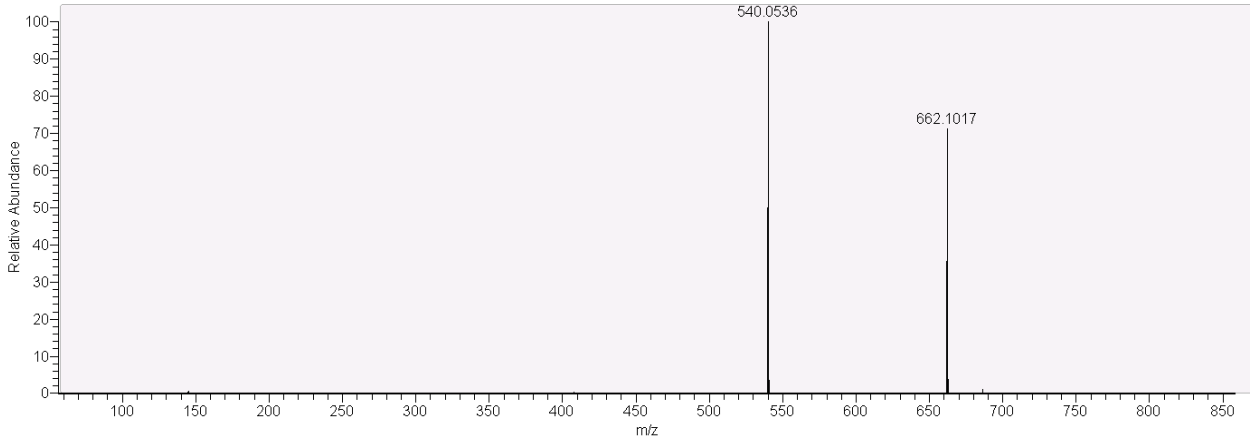


Figure S2. Observation of NAD, glutamate, and their situational adduct in negative mode.

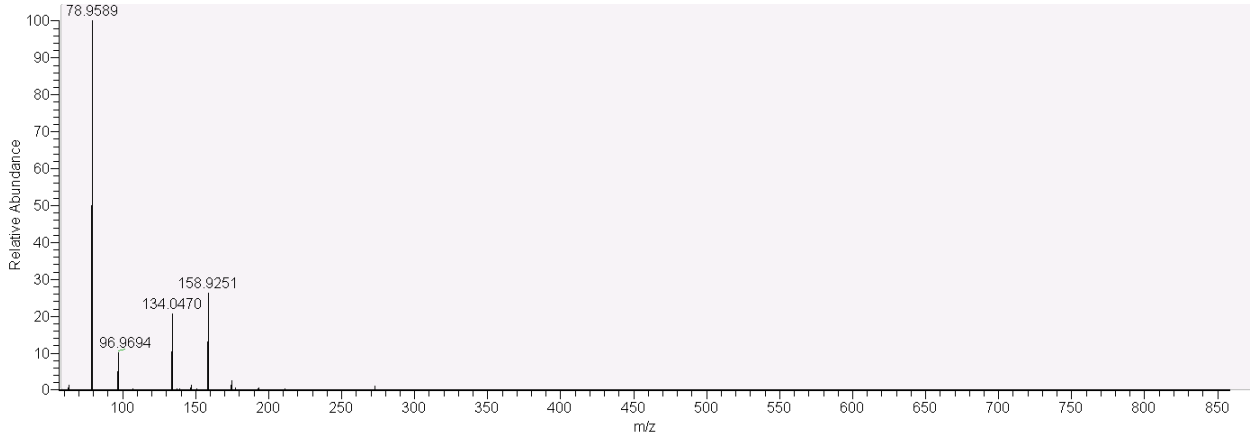
A. MS1 of glutamate and NAD (negative mode electrospray ionization).



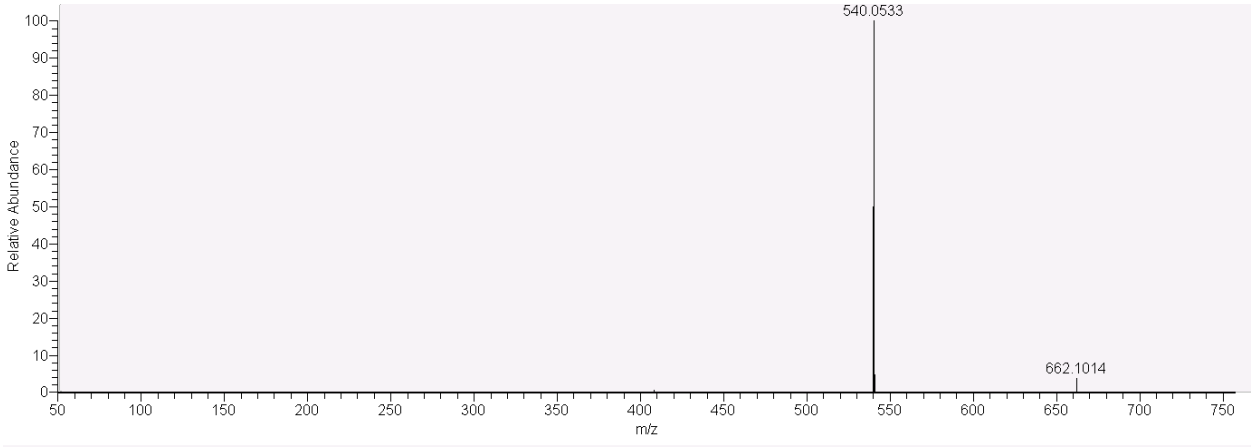
B. MS/MS of situational adduct $[Glutamate + NAD - H]^{1-}$, $[808.1710]^{1-}$. HCD = 10V.



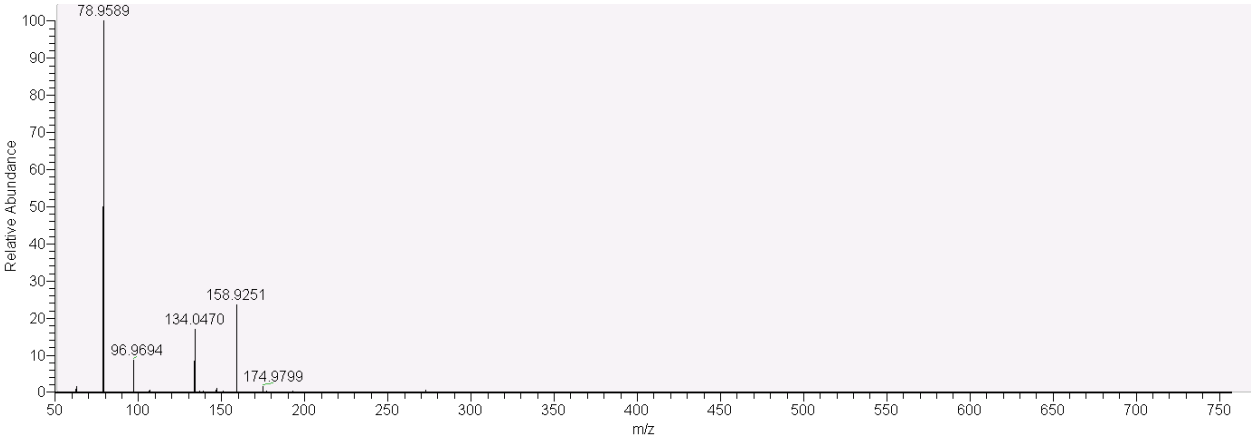
C. MS/MS of situational adduct $[Glutamate + NAD - H]^{1-}$, $[808.1710]^{1-}$. HCD = 60V.



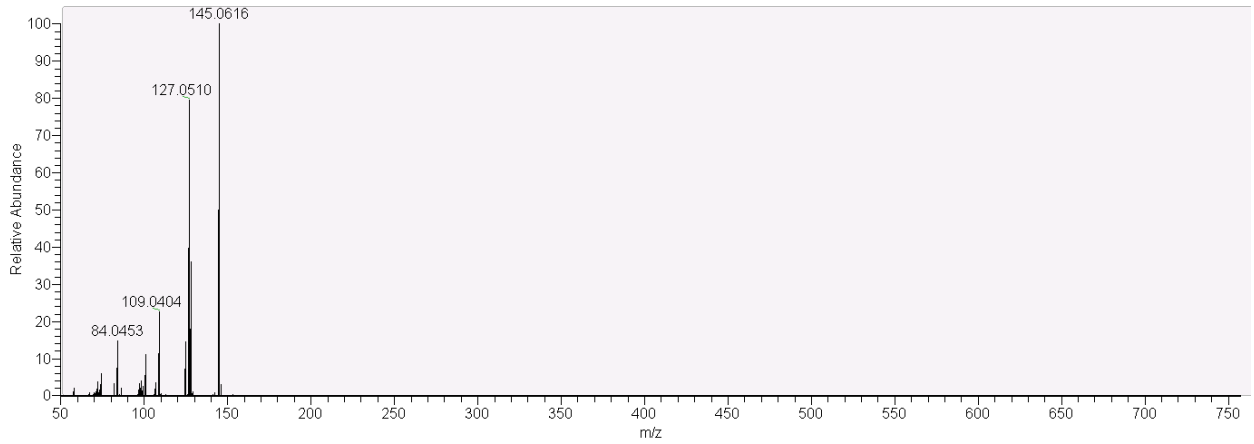
D. MS/MS of analyte $[NAD - H]^{1-}$, $[662.1021]^{1-}$. HCD = 10V.



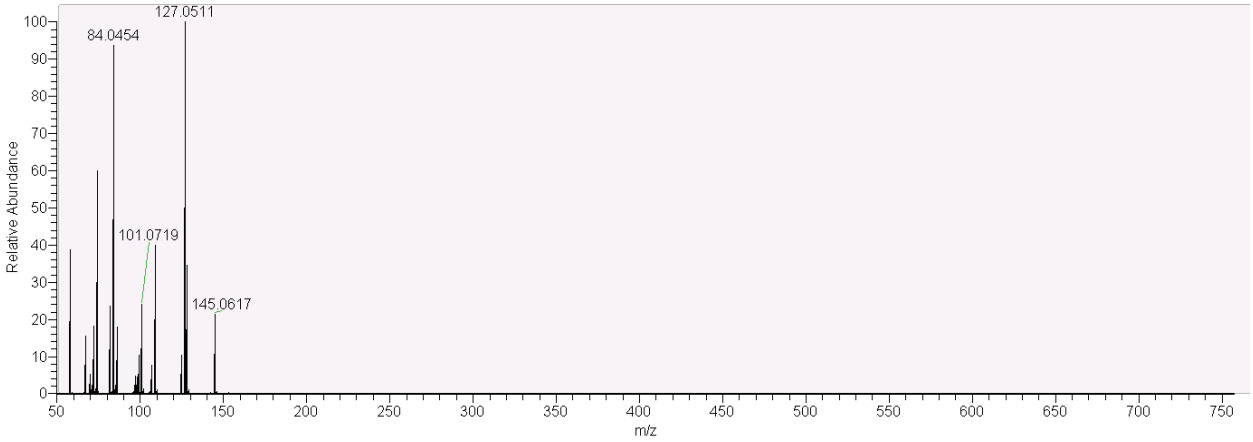
E. MS/MS of analyte $[NAD - H]^{1-}$, $[662.1021]^{1-}$. HCD = 60V.



F. MS/MS of analyte $[Glutamate - H]^{1-}$, $[145.0616]^{1-}$. HCD = 10V.



G. MS/MS of analyte $[Glutamate - H]^{1-}$, $[145.0616]^{1-}$. HCD = 10V.

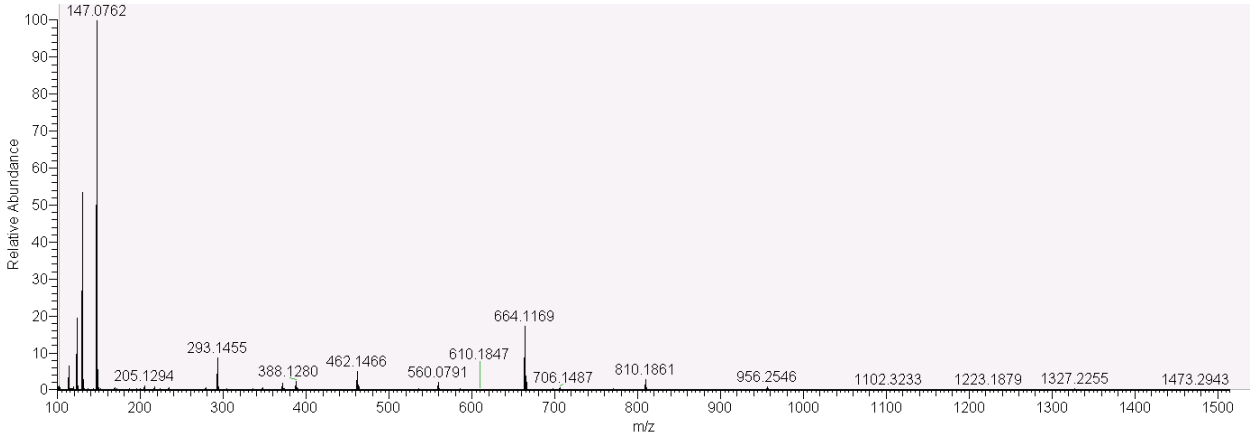

127.0511
84.0454
101.0719
145.0617
Relative Abundance
m/z
100
90
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750

52

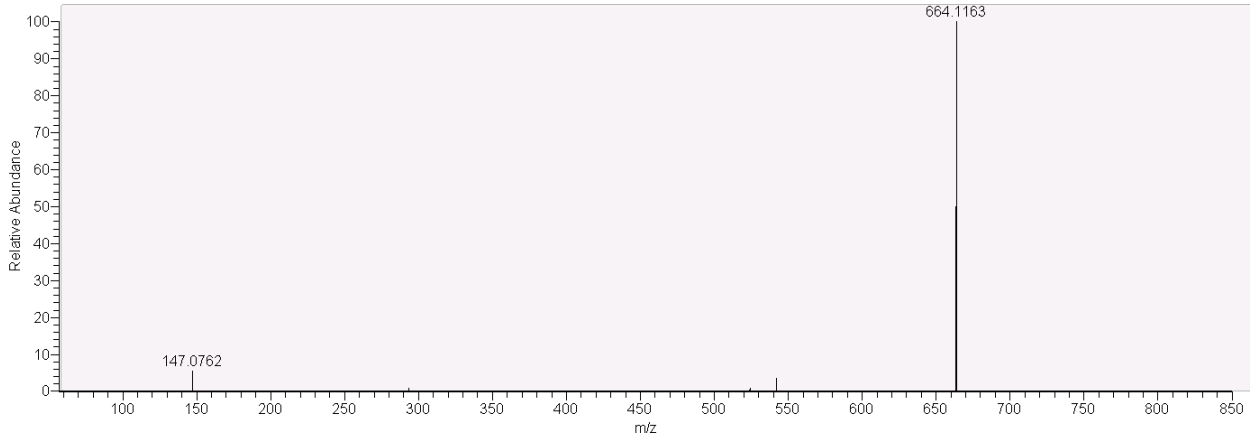
53

Figure S3. Observation of NAD, glutamate, and their situational adduct in positive mode.

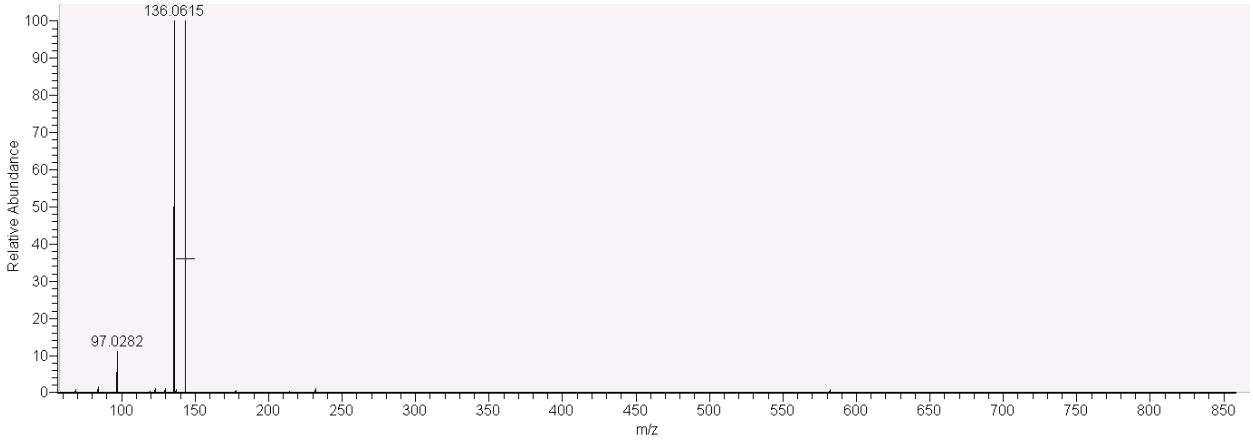
A. MS1 of glutamate and NAD (positive mode electrospray ionization).



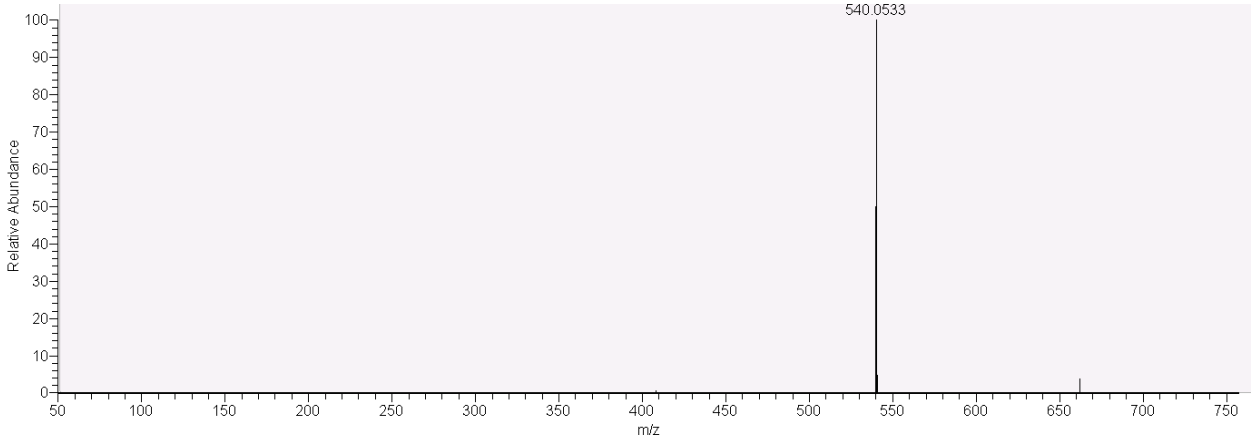
B. MS/MS of situational adduct [Glutamate + NAD + H]$^{1+}$, [810.1861]$^{1+}$. HCD = 10V.



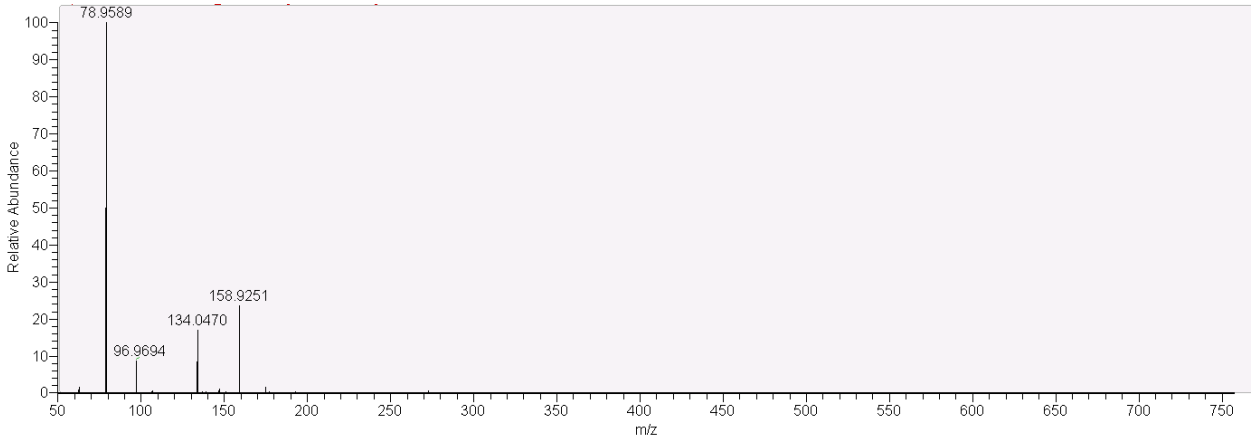
C. MS/MS of situational adduct [Glutamate + NAD + H]$^{1+}$, [810.1861]$^{1+}$. HCD = 60V.



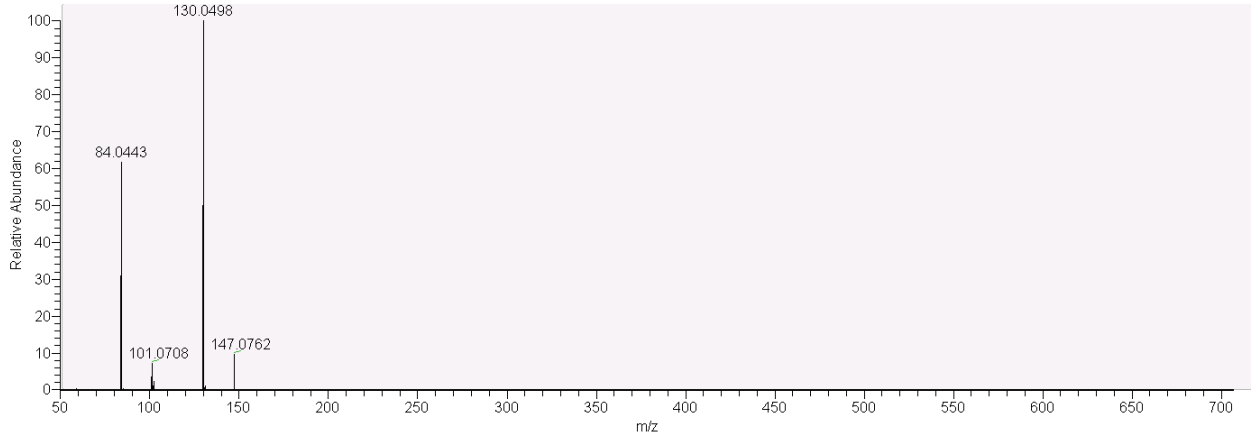
D. MS/MS of analyte [NAD + H]$^{1+}$, [664.1169]$^{1+}$. HCD = 10V.



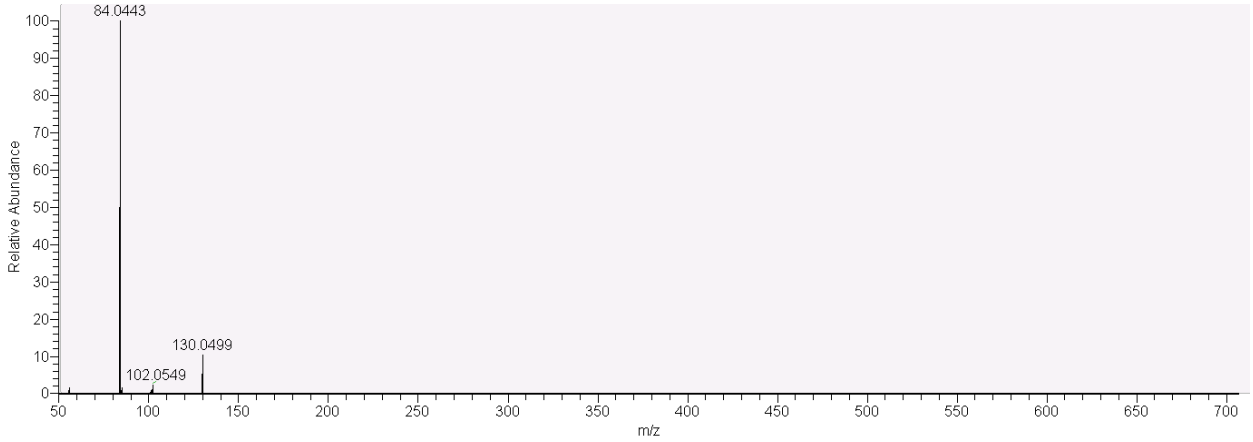
E. MS/MS of analyte $[NAD + H]^{1+}$, $[664.1169]^{1+}$. HCD = 60V.



C. MS/MS of analyte $[Glutamate + H]^{1+}$, $[147.0762]^{1+}$. HCD = 10V.



C. MS/MS of analyte $[Glutamate + H]^{1+}$, $[147.0762]^{1+}$. HCD = 60V.


84.0443
130.0499
102.0549
Relative Abundance
m/z
100
90
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700

69

70

Figure S4. Example of coeluting ions that form a multi-analyte dimer.

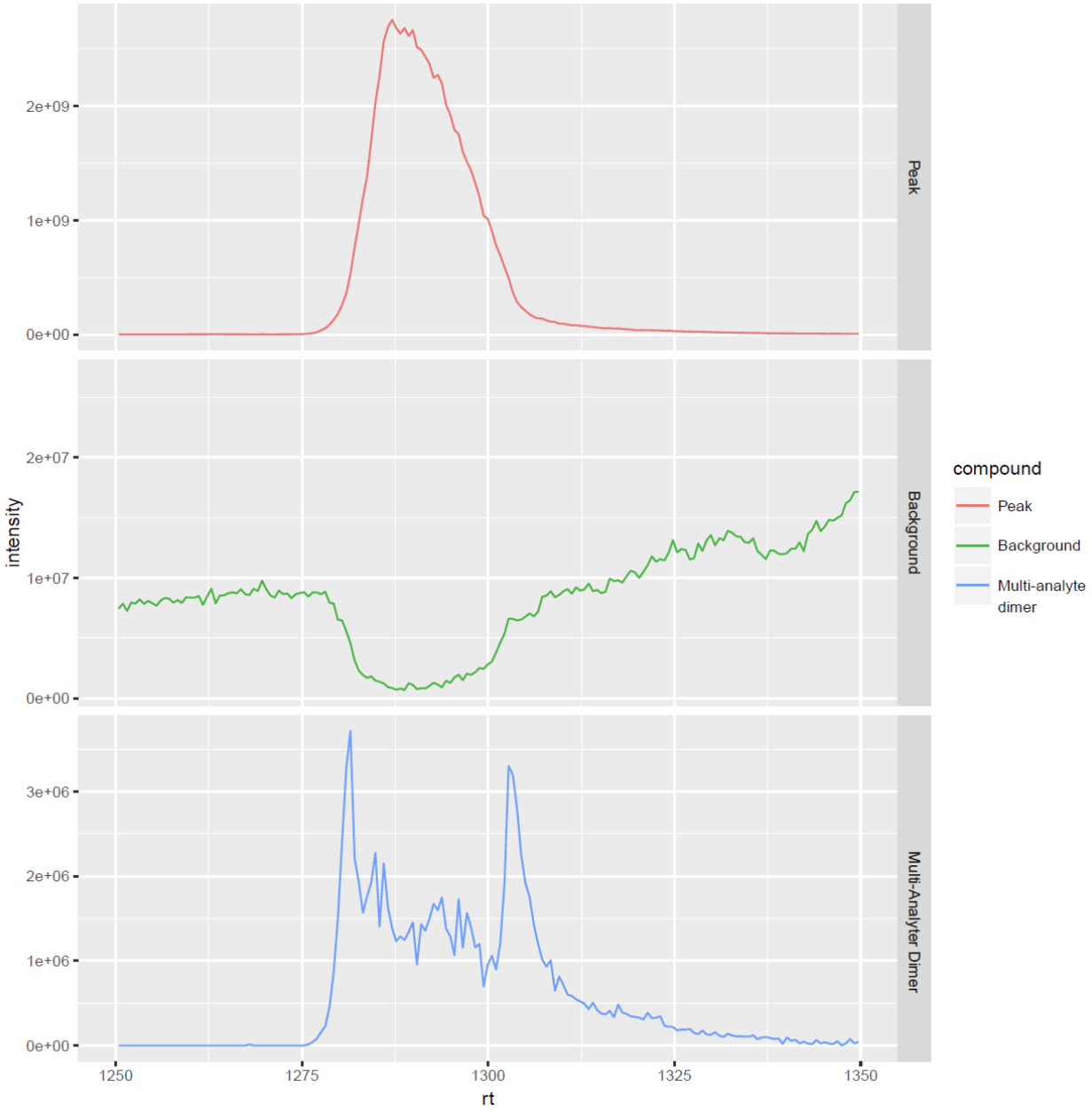


Figure S5. Screenshots from the creDBle database. (A) Partial list of credentialed features showing *m/z*, retention time, polarity, grouping, and intensity. (B) A credentialed features page showing the extracted ion chromatogram, credentialed isotopes, and fragmentation data.

A.

Features

Show 10 entries

Search: Show / hide columns

Feature ID	Component Group	Identity	+/-	m/z	RT (s)	Carbons	MS/MS	Intensity
cp.g2yd14bA	cc.gA6t3nq		+	754.3843	1103.3	16	0	1.0e+07
cp.g2yd14bq	cc.gA6tmm1		+	752.3669	1057.7	0	1	2.2e+05
cp.g2yd14br	cc.gA6tmm3		+	752.4848	829.9	0	0	1.2e+05
cp.g2yd14bs	cc.gA6tmmq		+	752.3904	1091.2	0	0	2.5e+06
cp.g2yd14bt			+	753.3996	1050.7	0	0	9.8e+04
cp.g2yd14bv	cc.gA6tmmy		+	755.2841	1361.0	43	0	6.5e+05

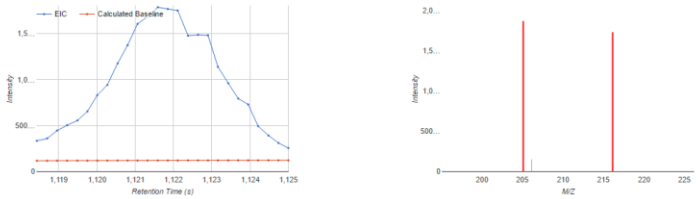
B.

cp.g2yd1mbx Details

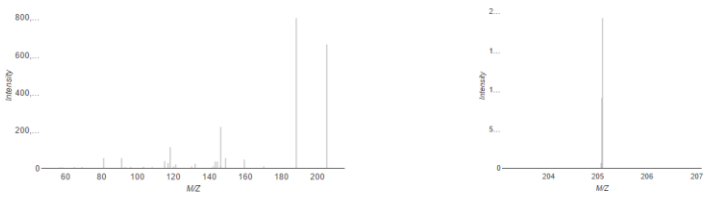
<< Previous Feature - Next Feature >>

Peak ID	m/z	RT	Intensity	Carbon Number	Polarity	Credentialed?	Peaks in Group	Experiment ID
cp.g2yd1mbx	710.4991	1121.6	1.7e+06	11	+	true	2	cm.h2

Spectral Data



Fragmentation Data



43.33V

Supplemental Methods

1 Materials

U-^{13}C-D-glucose was purchased from Cambridge Isotope Laboratories Inc. (Andover, MA). *E. coli* strain K12, MG1655 was purchased from ATCC (Manassas, VA). Lennox LB broth powder and 5x M9 salts were purchased from Sigma-Aldrich (St. Louis, MO). Cell culture was performed with ultrapure water provided by a Milli-Q system (Millipore). LC/MS grade, Burdick & Jackson brand water, acetonitrile, methanol, and isopropanol were purchased from Honeywell (Morris Plains, NJ). Cortecs T3 reversed phase UPLC columns and column guards were purchased from Waters Corporation (Milford, MA).

Generating Credentialed Samples

E. coli was grown in a rotary shaker at 37 ºC and 300 rpm as previously described.[27] A 100 mL volume of M9 minimal media was used with a glucose concentration of 2 g/L. Two cultures were grown in parallel, one using natural-abundance glucose and a second using U-^{13}C-glucose as the only carbon source. Cultures were grown to OD_{600} = 0.7, at which point they were harvested.

For harvest, flasks were removed from the shaker and placed on ice. The contents of each flask were pipetted into 50 mL conical tubes and centrifuged at 3200g and 4 ℃ for 10 minutes. The supernatant was decanted and remaining media was gently rinsed off the top of the pellet with 0.5 mL of water. Conical tubes were then placed in liquid nitrogen and lyophilized for 24 hours, or until dry. This powdered, credentialed *E. coli* standard was then extracted to generate samples for untargeted metabolomic analysis.

Several replicate extractions were performed in parallel by using a previously described method.[30] Briefly, five 2.5 mg samples of each ^{12}C and ^{13}C material were weighed out, while two empty tubes were included as extraction blanks. To these, 1,000 μL of 2:2:1

methanol:acetonitrile:water was added, followed by three freeze-thaw cycles with sonication and vortexing. After centrifugation, the supernatant was vacuum concentrated and reconstituted in 100 μL of 1:1 acetonitrile:water with internal standards. From these extracts, three samples were aliquoted for LC/MS analysis: natural-abundance extract, a mix of 1:1 natural-abundance extract and ^{13}C extract, and the blank extract.

Data Set Generation

The untargeted LC/MS data set was generated in positive polarity on a Q Exactive Plus mass spectrometer with a HESI II source coupled to a Dionex 3000RSLC. The data set was collected with the following settings: aux gas, 5; sheath gas, 35; sweep gas, 2; capillary temperature, 300 ºC; aux gas temperature, 200 ºC; spray voltage, 3.5 kV; needle diameter, 34 ga; s-lens, 75 V; mass range, 100–1500 Da; resolution 70,000; micro scans, 1; max injection time; 100 ms; automatic gain control target: 1e6. Reversed-phase chromatography was performed with the Waters Cortecs T3 (2.1mm x 50mm, 1.6um) column at a flow rate of 300 μL/min and a column temperature of 50 ºC. Solvents were: A, water + 5 mM ammonium acetate + 5 μM ammonium phosphate; B, 9:1 isopropanol:methanol + 5 mM ammonium acetate + 5 μm ammonium phosphate. An injection volume of 2 μL was used with a linear gradient of (minutes, %A): 0, 100; 28, 0; 30, 0; 30, 100; 35, 100.

Chromatographic features were detected. Mass traces were retained if they were longer than 10 scans, excluding missing peaks. Baselines for each mass trace were calculated by using the iterative restricted least squares method from the baseline R package. Model based peak detection was performed by using the skew normal distribution as a model peak distribution. This process resulted in a set of features detected in each replicate run. Features were grouped by mass and retention time using a density based method. Retention time drift and mass drift were corrected by fitting a loess curve of degree 2 to the distance from the mean value of each group against the mean retention time of each group. An outline of the workflow

can be found on GitHub at https://github.com/nathaniel-mahieu/metabolomic-feature-reduction-2017.

Subtle variations from run to run cause many features to be integrated differently and sometimes not integrated in each file. Further, closely eluting peaks often lead to incorrectly grouped features. To resolve these missing values, refine the individual data sets, and get a set of detected peaks consistent with all replicate runs, we applied the Warpgroup algorithm.[26] Warpgroup is available at https://github.com/nathaniel-mahieu/warpgroup. Warpgroup takes as input the raw data and each file's detected features combining them to output a set of consensus features. Parameters: sc.aligned.lim, 9; pct.pad, 0.1; min.peaks, 3. Of the detected peaks we retained only features with a signal-to-noise ratio >5 and a coefficient of variation <0.5 after Warpgrouping. This resulted in 25,230 "high-quality" features in our representative data set.

Annotation Notes

Features can belong to only one group. Most group assignements are clear (e.g., a sodium adduct of glutamate belongs in the glutamate group). The only ambiguous group assignments are when two different metabolites dimerize (e.g., a dimer between glutamate and NAD). In this case, our assignment of the dimer feature to either the glutamate or NAD group is arbitrary. We note that the number of features in a group does not affect our overarching goal to count the number of unique metabolites present in the data set.

Relationship graphs are nonlinear, span many peaks, and specify only transformations of masses. M+H, M+Na, and M+K form a fully connected graph. Determination of the original monoisotopic mass is a related but distinct question not answerable with current methods.