**Supporting Information**

# AutoTuner:
# High Fidelity and Robust Parameter Selection for Metabolomics Data Processing

Craig McLean[1,2,+] and Elizabeth B. Kujawinski[1]

[1] Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA
[2] MIT/WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering, Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA
[+] Corresponding Author: crmclean@mit.edu

**Table of Contents:**

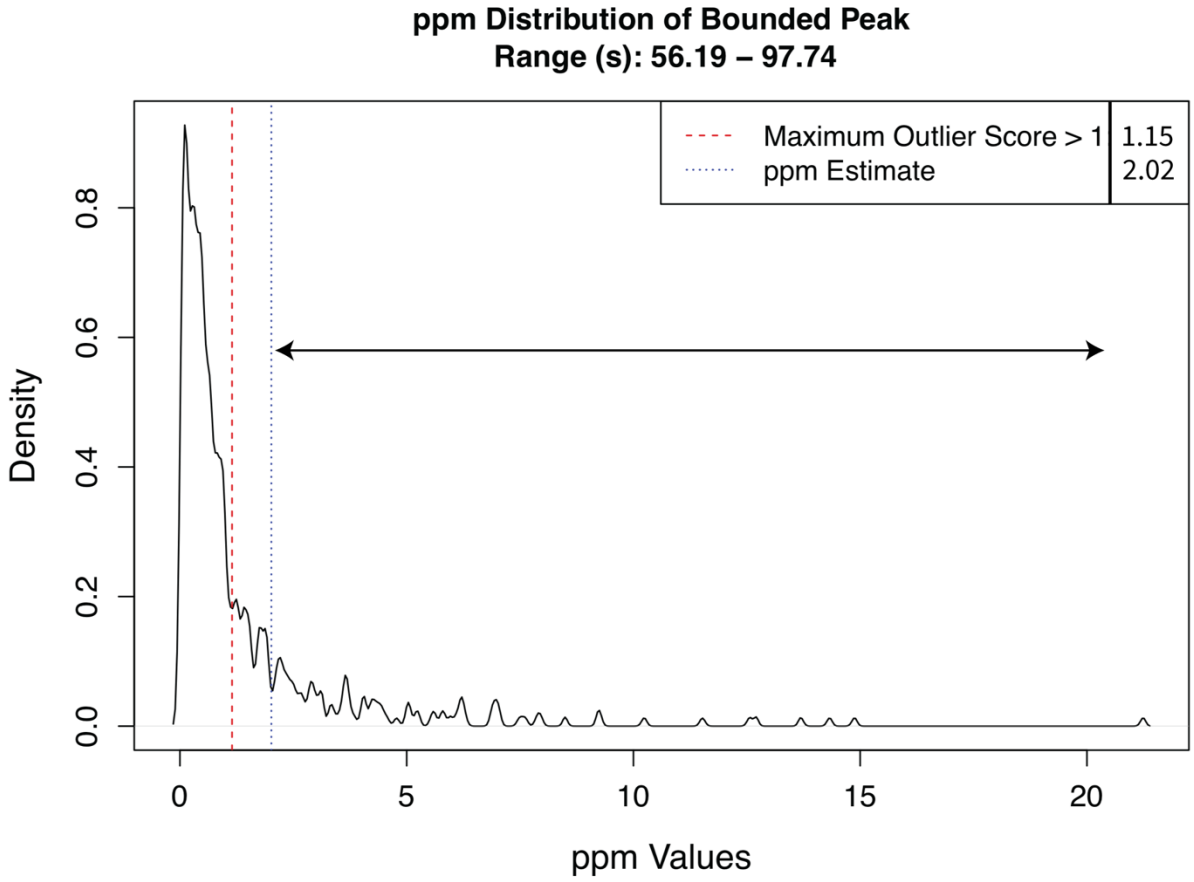


**Figure S1.** Example of AutoTuner-generated ppm error distribution. Such plots are returned by the algorithm to check quality of estimates. Red line represents the maximum ppm error value with an outlier score greater than 1 (see equation 3). In this example, a ppm error value of 1.15 meets this criterion (see legend). Blue line represents the ***ppm*** error parameter estimate described in equation 4, or 2.02 in this example (see legend). The "Range" subtitle represents the original chromatographic bounds of the TIC peak used to obtain estimates. The peaks under the arrow are assumed to originate from ppm values calculated from random associations of noise rather than from true features.

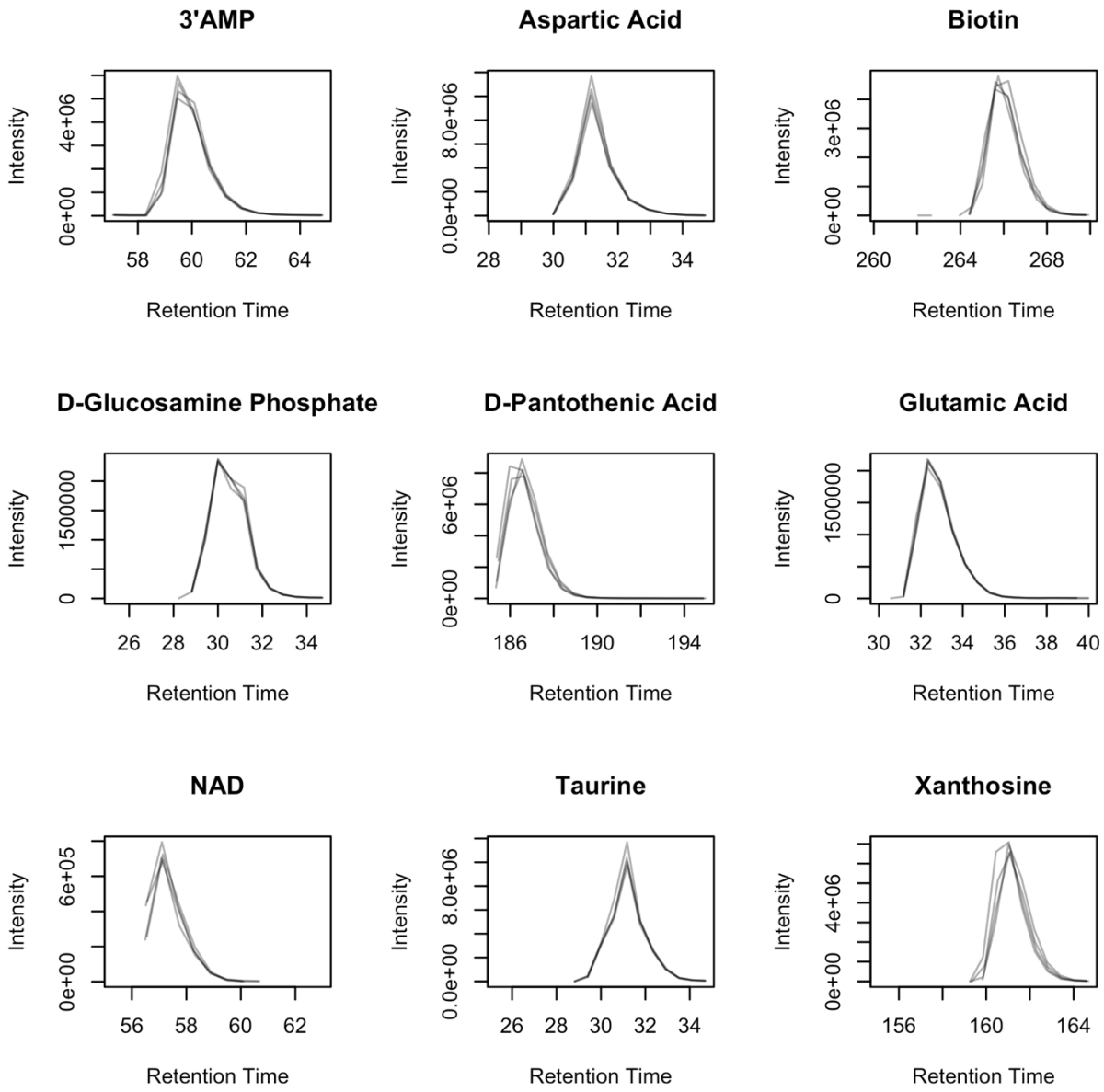


**Figure S2.** Example EIC peaks of standards not detected within feature table generated with IPO-derived parameters. The lines represent individual standard samples. 3'AMP = 3'-adenosine monophosphate; NAD = ß-nicotinamide adenine dinucleotide

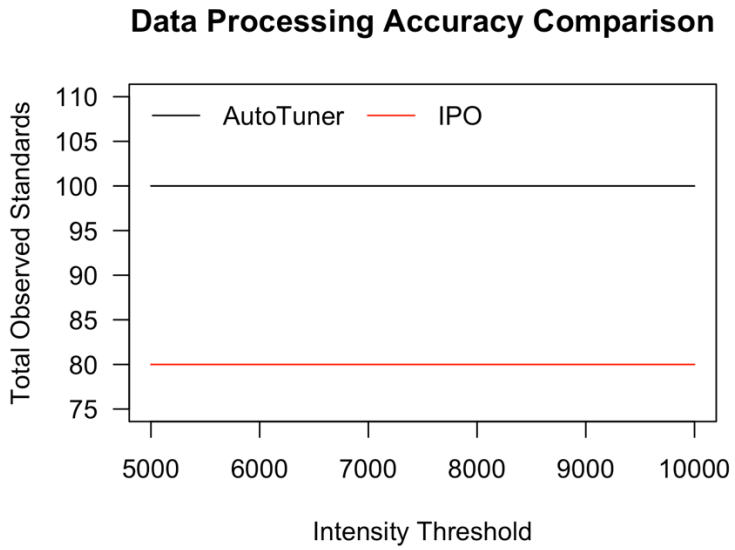


**Figure S3.** Impact of feature intensity threshold on standard detection. Intensity threshold varied by 5000 from 5000 until 10000. Lines indicate the number of detected standards from feature tables generated with IPO- and AutoTuner-derived parameters. The minimum intensity value observed across standards was measured at 74804.84.

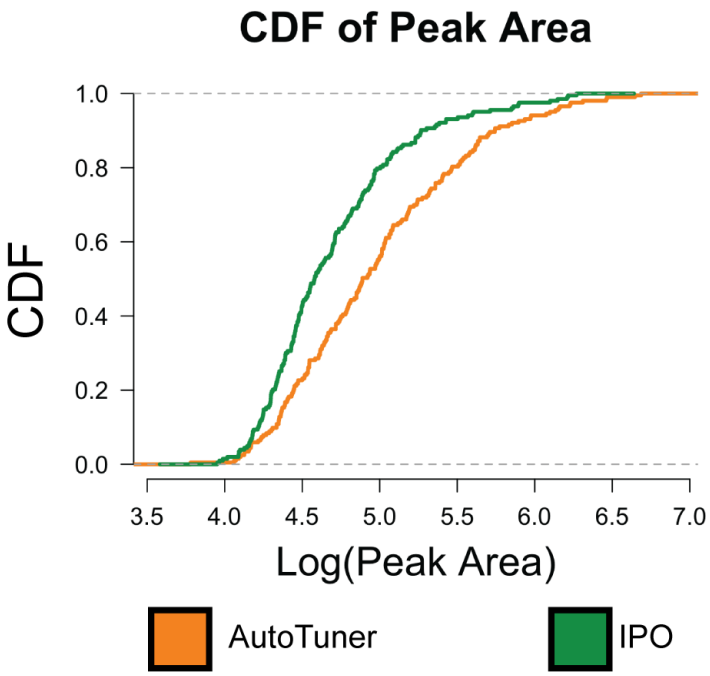


**Figure S4.** Positive ion mode data empirical cumulative distribution functions (CDF) comparison of peak area from EICs of features uniquely identified within feature tables generated with AutoTuner- and IPO-derived parameters. The curves were significantly different from one another (KS-test, Area: $p < 10^{-6}$; n = 203).

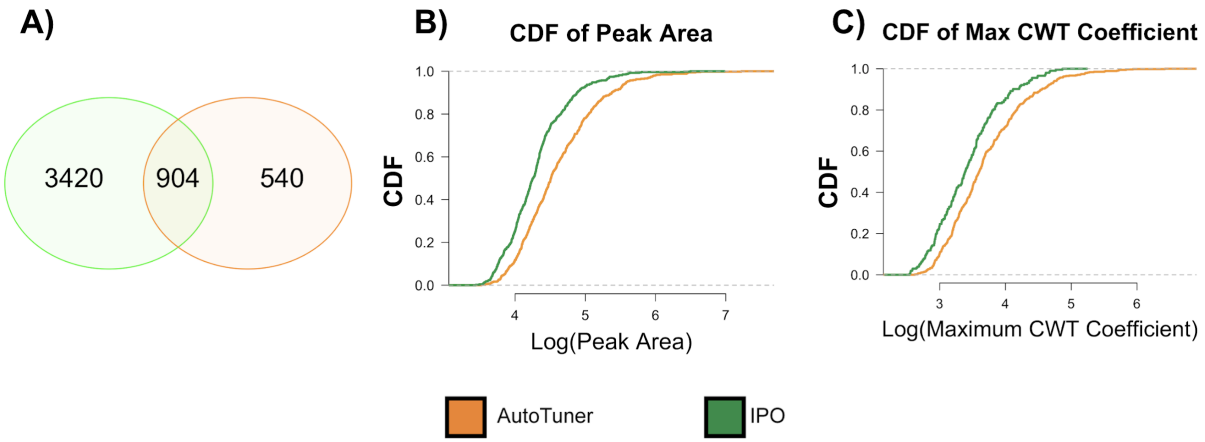


**Figure S5.** Negative ion mode data comparison of feature tables based on AutoTuner- and IPO-derived parameters on the culture dataset. A) portrays the overlap in the number of m/z-rt features generated by both methods. Features with an error of 5 ppm and retention time error of 20 seconds are placed in the intersect. B and C compare the differences in structural properties for the (B) peak area and (C) maximum continuous wavelet transform coefficient (CWT) between peaks detected only within AutoTuner or IPO. Both curves are empirical cumulative distribution functions (CDF) of the calculated metrics. An empirical cumulative distribution function is a non-parametric estimator of the underlying CDF of a random variable. In this case, the random variable is the set of calculated values for the AutoTuner- and IPO-specific features. CDFs for each metric were significantly different from one another (KS-test, Area: $p < 10^{-14}$; CWT: $p < 10^{-8}$, $n = 540$), similar to positive ion mode data.

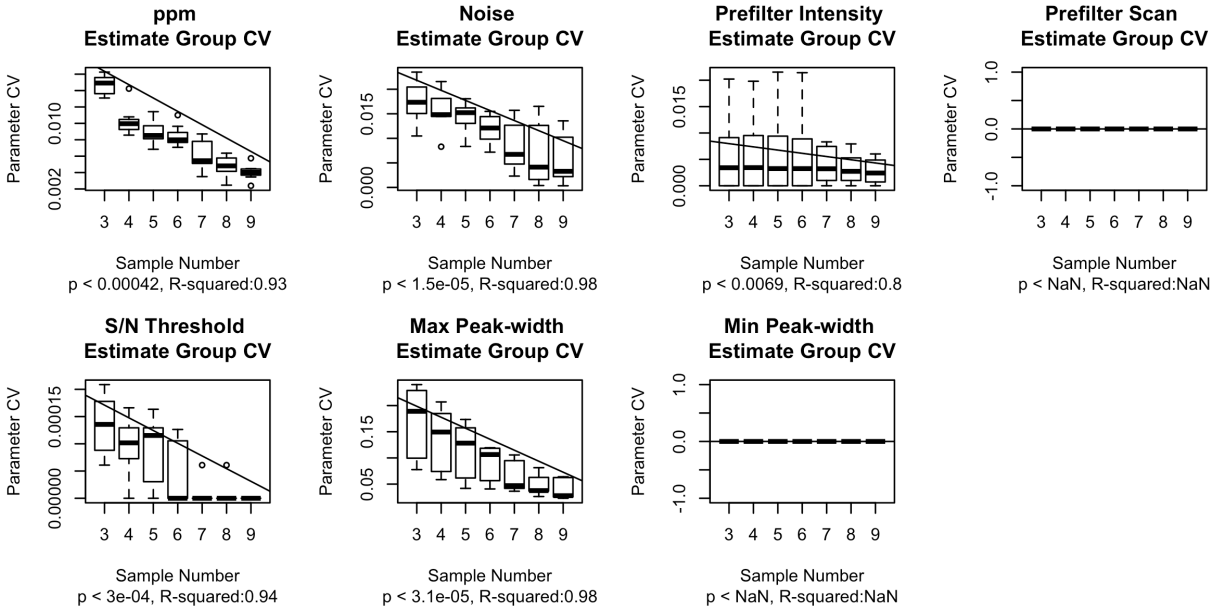


**Figure S6.** The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on negative mode community data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and $R^2$ statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

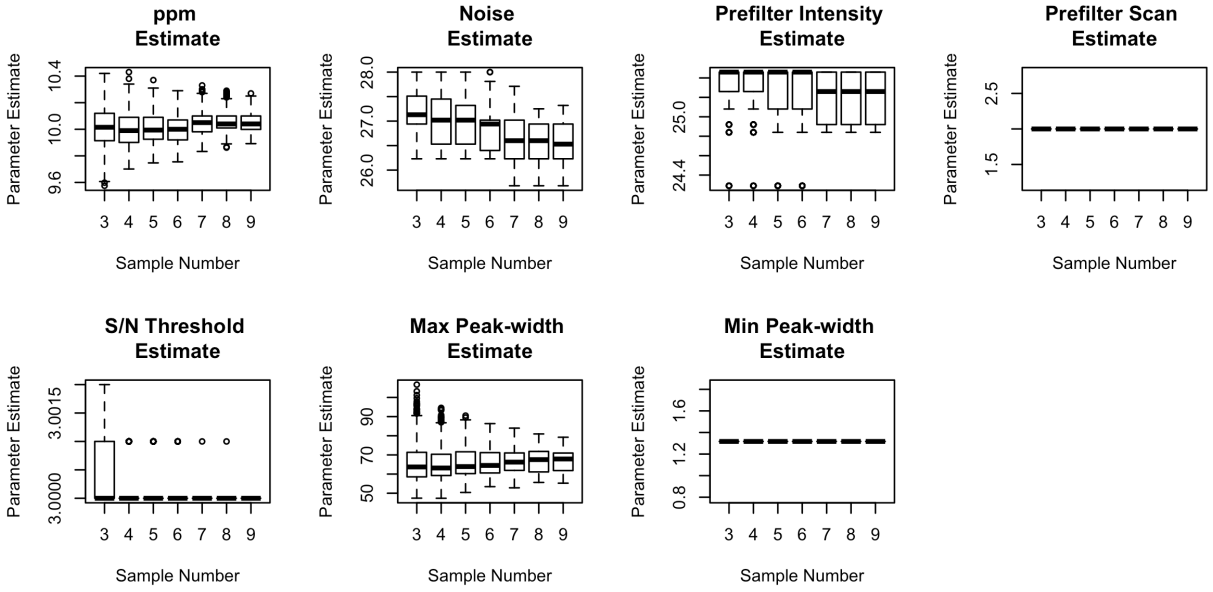


**Figure S7.** The parameters estimated in the Monte Carlo analysis on negative mode community data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

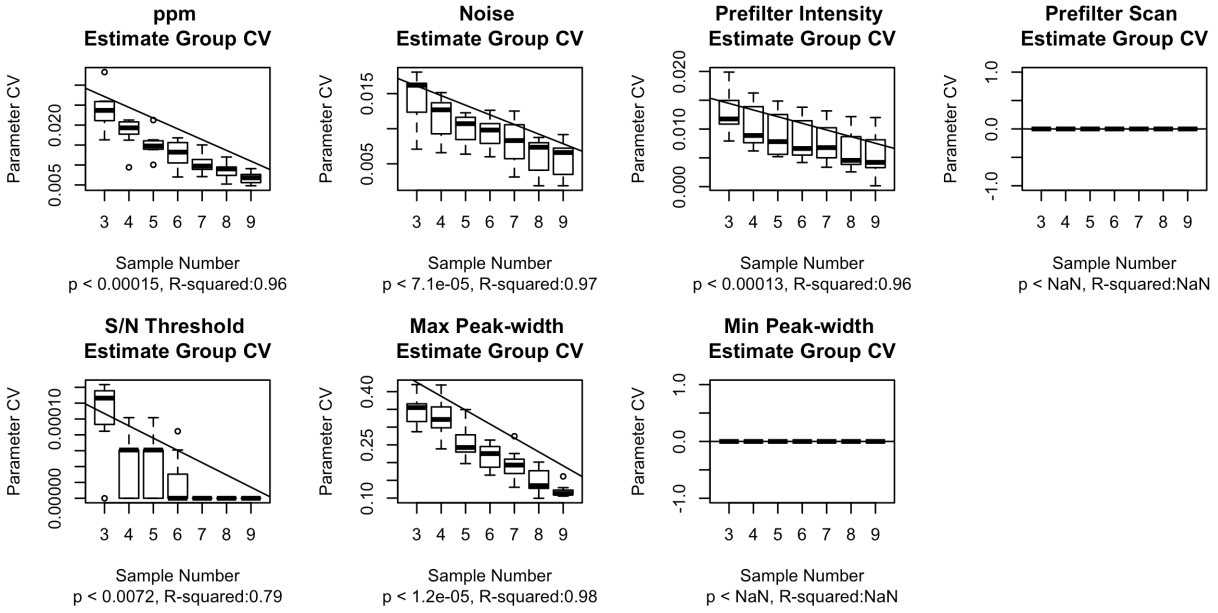


**Figure S8.** The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on positive mode community data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and $R^2$ statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

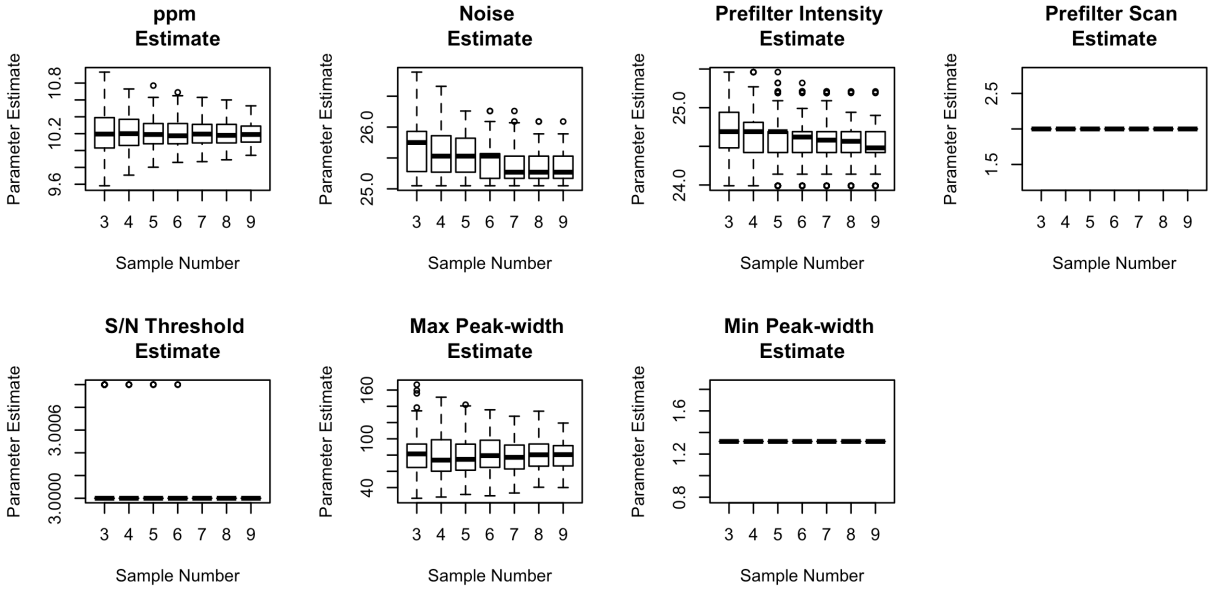


**Figure S9.** The parameters estimated in the Monte Carlo analysis on positive mode community data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

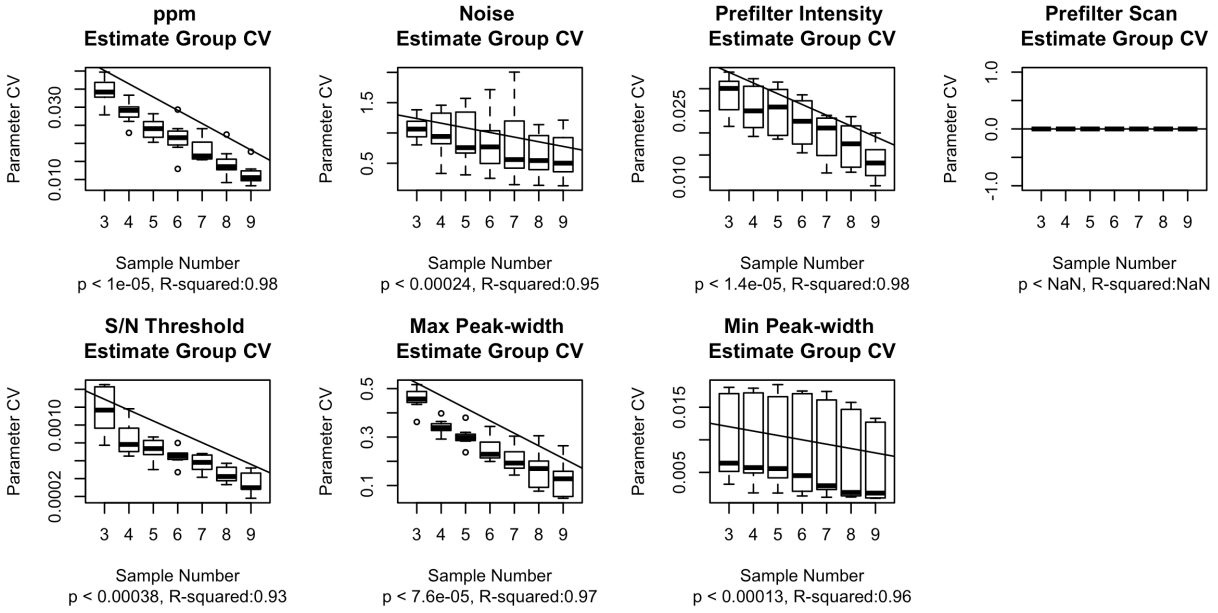


**Figure S10.** The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on negative mode culture data. Each plot denotes the calculated CV values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and $R^2$ statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

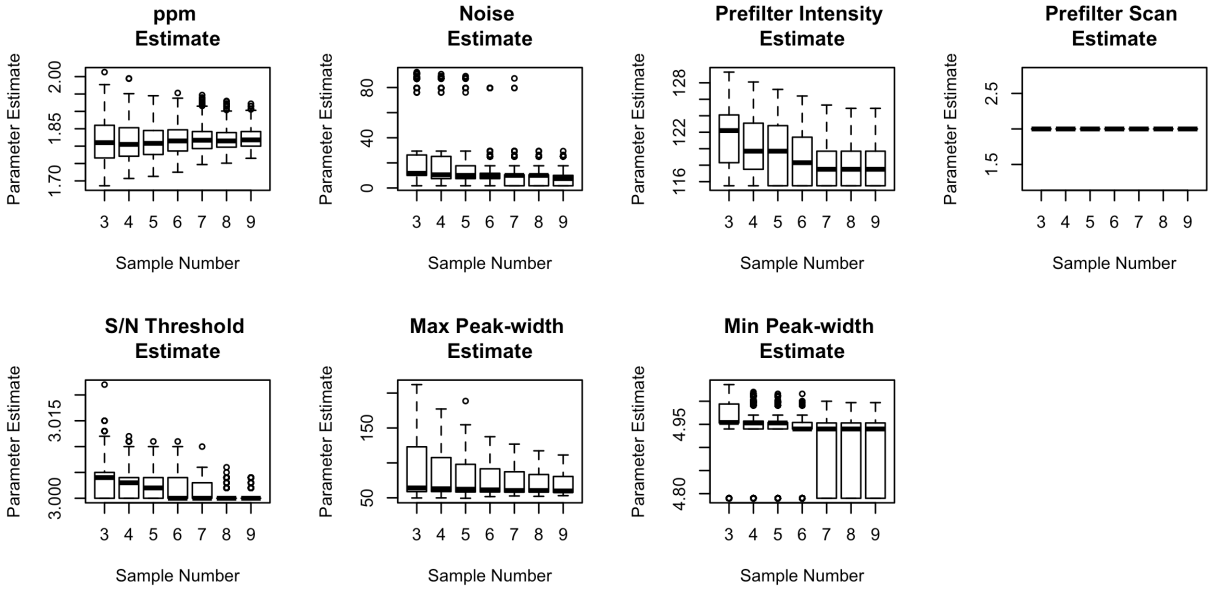


**Figure S11.** The parameters estimated in the Monte Carlo analysis on negative mode culture data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

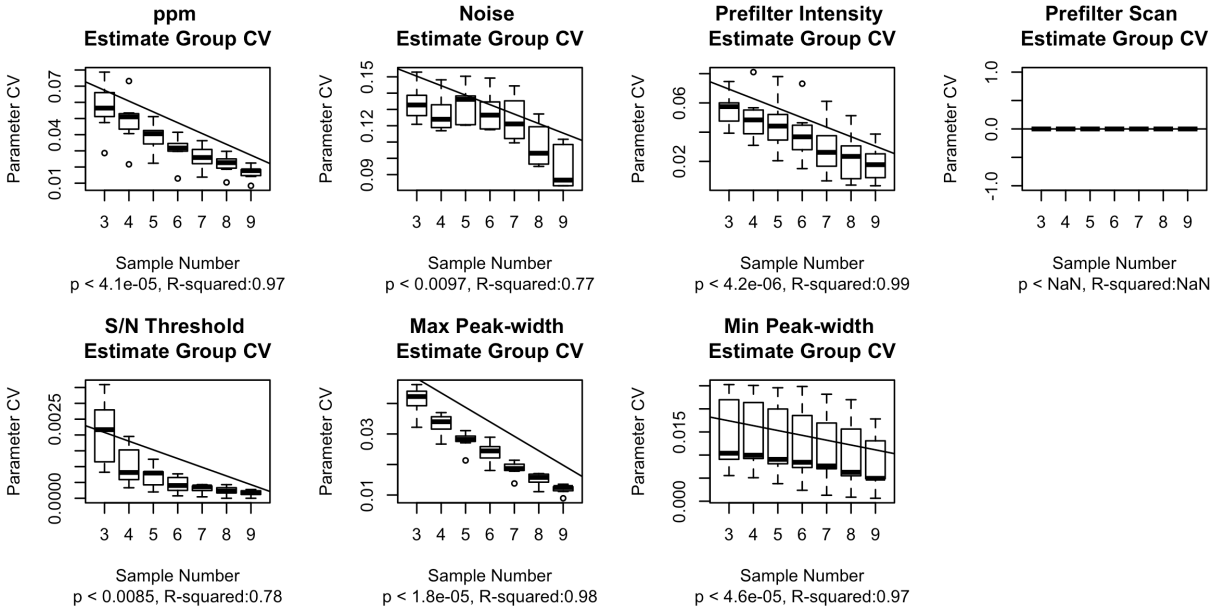


**Figure S12.** The coefficient of variation (CV) for groups of parameters estimated in the Monte Carlo analysis on positive mode culture data. Each plot denotes the calculated values for each unique parameter. The x-axis describes the number of samples used to generate estimates, while the y-axis describes the CV of the estimates from each group of 11 randomly selected samples. P-value and $R^2$ statistics are derived from linear regressions of data (n = 49). (NaN = not a number).

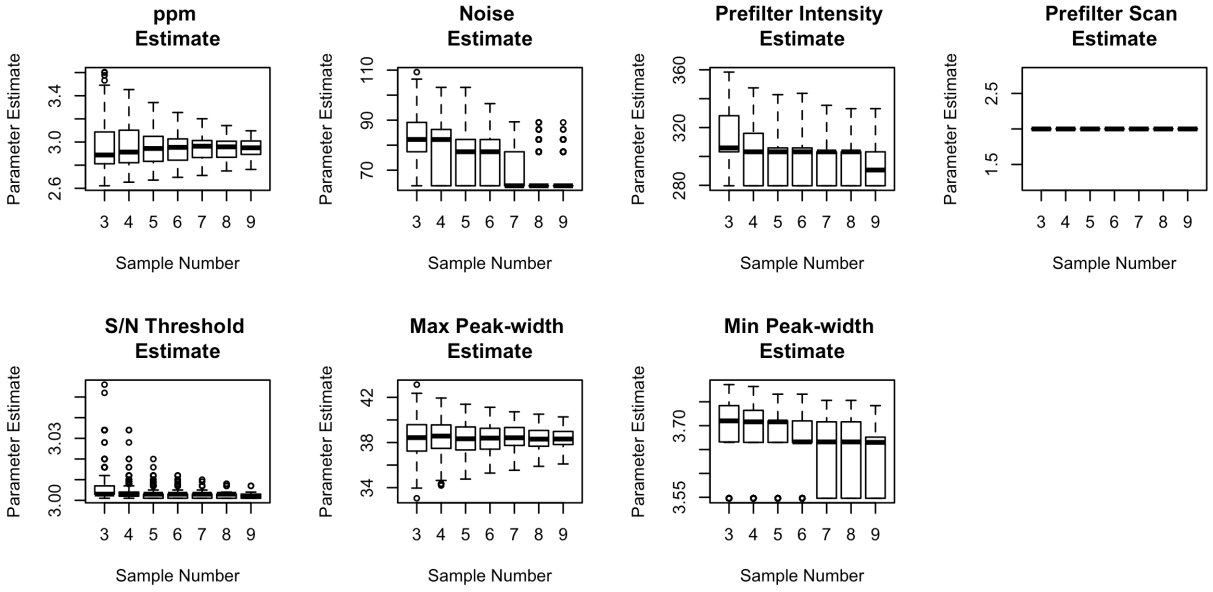


**Figure S13.** The parameters estimated in the Monte Carlo analysis on positive mode culture data. Each plot denotes the calculated parameter estimate values for each unique parameter across 385 runs of AutoTuner. The x-axis describes the number of samples used to generate estimates, while the y-axis portrays the determined 55 parameter estimates within each n-sample subset (n = 3-9).

**Table S1.** Standards used to validate AutoTuner accuracy. These compounds are common targets of metabolism and are commonly detected within untargeted metabolomics experiments. Compounds detected in both ionization modes are separated by "|" in the order they were presented in the "Ionization Mode" column.

| Compound | Ionization Mode | In AutoTuner | In IPO | m/z | Retention Time (s) |
|---|---|---|---|---|---|
| 3-methyl-2-oxopentanoic acid | NEG | TRUE | TRUE | 129.061\|NA | 237.7 |
| 3-methyl-2-oxobutanoic acid | NEG | TRUE | TRUE | 115.05\|NA | 149.2 |
| 4-aminobenzoic acid | POS | TRUE | TRUE | NA\|138.043 | 202.2 |
| 4-hydroxybenzoic acid | NEG | TRUE | TRUE | 137.028\|NA | 235.4 |
| 4-methyl-2-oxopentanoic acid | NEG | TRUE | TRUE | 129.061\|NA | 255.3 |
| adenosine 5'-monophosphate (5'AMP) | NEG\|POS | TRUE\|TRUE | FALSE\|TRUE | 346.039\|348.054 | 53.0 |
| adenosine 3'-monophosphate (3'AMP) | NEG\|POS | TRUE\|TRUE | FALSE\|TRUE | 346.039\|348.054 | 59.5 |
| 6-phosphogluconic acid | NEG | TRUE | TRUE | 275.002\|NA | 32.9 |
| acetyl taurine | NEG | TRUE | TRUE | 166.017\|NA | 43.6 |
| adenine | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 134.053\|136.063 | 52.4 |
| adenosine | POS | TRUE | TRUE | NA\|268.091 | 110.6 |
| alpha-ketoglutaric acid | NEG | TRUE | TRUE | 145.039\|NA | 53.0 |
| 4-amino-5-aminomethyl-2-methylpyrimidine (AmMP) | POS | TRUE | TRUE | NA\|139.1 | 27.1 |
| arginine | POS | TRUE | FALSE | NA\|175.103 | 30.1 |
| aspartic acid | NEG\|POS | TRUE\|TRUE | FALSE\|TRUE | 132.025\|134.055 | 31.2 |
| biotin | NEG\|POS | TRUE\|TRUE | FALSE\|FALSE | 243.069\|245.073 | 266.2 |

| | | | | | |
|---|---|---|---|---|---|
| caffeine | POS | TRUE | TRUE | NA\|195.069 | 248.1 |
| citric acid | NEG | TRUE | TRUE | 191.005\|NA | 59.5 |
| cytosine | POS | TRUE | TRUE | NA\|112.054 | 34.2 |
| desthiobiotin | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 213.109\|215.117 | 285.7 |
| glucosamine phosphate | NEG | TRUE | FALSE | 258.026\|NA | 30.6 |
| pantothenic acid | NEG\|POS | TRUE\|TRUE | FALSE\|TRUE | 218.094\|220.108 | 186.6 |
| ribose 5-phosphate | NEG | TRUE | TRUE | 229.006\|NA | 32.3 |
| 3-phosphoglyceric acid | NEG | TRUE | TRUE | 184.986\|NA | 35.9 |
| diacetylchitobiose | POS | TRUE | TRUE | NA\|425.132 | 44.2 |
| dihydroxy acetone phosphate | NEG | TRUE | TRUE | 168.984\|NA | 32.3 |
| dimethylsulfoniopropionate (DMSP) | POS | TRUE | TRUE | NA\|135.052 | 31.9 |
| ectoine | POS | TRUE | TRUE | NA\|143.14 | 38.3 |
| folic acid | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 440.101\|442.121 | 230.3 |
| fosfomycin | NEG | TRUE | TRUE | 137.011\|NA | 37.7 |
| fumarate | NEG | TRUE | TRUE | 115.007\|NA | 71.0 |
| gamma-aminobutyric acid (GABA) | POS | TRUE | TRUE | NA\|104.087 | 32.4 |
| glucose 6-phosphate | NEG | TRUE | TRUE | 259.004\|NA | 31.2 |
| glutamic acid | NEG | TRUE | FALSE | 146.047\|NA | 32.9 |
| glutamine | POS | TRUE | FALSE | NA\|147.073 | 31.3 |
| glycine betaine | POS | TRUE | TRUE | NA\|118.08 | 34.8 |
| glyphosate | NEG | TRUE | TRUE | 168.061\|NA | 31.7 |
| guanine | POS | TRUE | TRUE | NA\|152.064 | 53.0 |
| guanosine | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 282.06\|284.099 | 137.6 |
| 4-methyl-5-thiazoleethanol (HET) | POS | TRUE | TRUE | NA\|144.057 | 178.1 |

| | | | | | |
|---|---|---|---|---|---|
| (4-amino-2-methyl-5-pyrimidinyl)methanol (HMP) | POS | TRUE | TRUE | NA\|140.084 | 46.5 |
| c 3-acetic acid | POS | TRUE | TRUE | NA\|176.065 | 318.6 |
| inosine | NEG | TRUE | TRUE | 267.061\|NA | 138.5 |
| inosine 5'-monophosphate | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 347.022\|349.037 | 57.1 |
| isethionic acid | NEG | TRUE | TRUE | 125.055\|NA | 34.1 |
| citrulline | POS | TRUE | TRUE | NA\|176.089 | 33.0 |
| glutathione | POS | TRUE | TRUE | NA\|308.053 | 77.7 |
| glutathione oxidized | POS | TRUE | TRUE | NA\|613.161 | 77.7 |
| isoleucine | POS | TRUE | TRUE | NA\|132.092 | 87.3 |
| kynurenine | POS | TRUE | TRUE | NA\|209.12 | 159.9 |
| leucine | POS | TRUE | TRUE | NA\|132.091 | 82.9 |
| phenylalanine | POS | TRUE | TRUE | NA\|166.079 | 166.3 |
| tryptophan | POS | TRUE | TRUE | NA\|205.084 | 213.9 |
| tyrosine | POS | TRUE | TRUE | NA\|182.105 | 81.5 |
| methionine | POS | TRUE | FALSE | NA\|150.052 | 57.1 |
| 5'methylthioadenosine (MTA) | POS | TRUE | TRUE | NA\|298.081 | 209.8 |
| muramic acid | NEG | TRUE | TRUE | 250.086\|NA | 40.0 |
| N-acetyl d-glucosamine | POS | TRUE | TRUE | NA\|222.077 | 37.2 |
| N-acetyl l-glutamic acid | NEG | TRUE | TRUE | 188.054\|NA | 70.1 |
| N-acetylmuramic acid | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 292.085\|294.121 | 109.6 |
| ß-nicotinamide adenine dinucleotide (NAD) | NEG\|POS | TRUE\|TRUE | FALSE\|FALSE | 662.041\|664.078 | 57.1 |
| ß-nicotinamide adenine dinucleotide phosphate (NADP) | NEG | TRUE | TRUE | 742.011\|NA | 53.9 |

| | | | | | |
|---|---|---|---|---|---|
| ornithine | POS | TRUE | TRUE | NA\|133.098 | 27.7 |
| orotic acid | NEG | TRUE | TRUE | 155.004\|NA | 50.1 |
| phosphoenolpyruvate | NEG | TRUE | TRUE | 166.970\|NA | 37.7 |
| proline | POS | TRUE | TRUE | NA\|116.076 | 32.2 |
| pyridoxine | POS | TRUE | TRUE | NA\|170.079 | 61.2 |
| riboflavin | POS | TRUE | FALSE | NA\|377.100 | 262.4 |
| S-(1,2-dicarboxyethyl)glutathione | POS | TRUE | TRUE | NA\|424.121 | 65.9 |
| S-(5'-adenosyl) -L-homocysteine (SAH) | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 383.054\|385.062 | 78.7 |
| S-adenosyl-l-methionine (SAM) | POS | TRUE | FALSE | NA\|399.200 | 31.3 |
| serine | POS | TRUE | FALSE | NA\|106.052 | 30.7 |
| sn-glycerol 3-phosphate | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 170.999\|173.004 | 32.3 |
| succinic acid | NEG | TRUE | TRUE | 117.022\|NA | 76.6 |
| syringic acid | NEG | TRUE | TRUE | 197.030\|NA | 266.2 |
| taurine | NEG | TRUE | FALSE | 124.012\|NA | 43.6 |
| thiamine monophosphate | POS | FALSE | FALSE | NA\|345.060 | NA |
| threonine | POS | TRUE | TRUE | NA\|120.069 | 31.9 |
| thymidine | NEG | TRUE | TRUE | 241.074\|NA | 173.8 |
| triacetylchitotriose | POS | TRUE | TRUE | NA\|628.269 | 53.0 |
| uracil | POS | TRUE | TRUE | NA\|113.051 | 114.0 |
| uridine 5'-monophosphate | POS | TRUE | TRUE | NA\|325.031 | 51.2 |
| valine | POS | TRUE | TRUE | NA\|118.091 | 34.8 |
| xanthine | NEG\|POS | TRUE\|TRUE | TRUE\|TRUE | 151.017\|153.045 | 161.0 |
| xanthosine | NEG\|POS | TRUE\|TRUE | FALSE\|TRUE | 283.053\|285.084 | 161.0 |

**Table S2.** Parameters used to process data. We rounded the values returned by AutoTuner and IPO at the tenths place. Each column aside from the "Dataset" and "Method" represent XCMS parameters described in Table 1. The community dataset is not mentioned here, as no comparison between IPO- and AutoTuner-parametrized feature tables was performed. The same standard set of parameters were used for density grouping and loess spline retention time correction. XCMS function syntax is described in parentheses. For the first run of density grouping (group.density): group difference =10, minfrac = 0, minsamp = 1, mzwid = 0.001. For the second run of density grouping after retention time correction (group.density):, group difference = 5, minfrac = 0.5, minsamp = 1, mzwid = 0.001. For loess spline retention time correction (retcor.peakgroups): span = 0.5.

| **Dataset** | **Method** | ***Maximum Peak-width*** | ***Minimum Peak-width*** | ***ppm*** | ***Noise*** | ***Prefilter Intensity*** | ***Scan Count*** | ***S/N Threshold*** |
|---|---|---|---|---|---|---|---|---|
| Pos Standards | IPO | 26.0 | 12.0 | 6.2 | 250.0 | 100.0 | 3.6 | 10 |
| Pos Standards | AutoTuner | 29.3 | 5.7 | 4.0 | 436.8 | 1421.3 | 2.0 | 6 |
| Pos Culture | IPO | 48.0 | 18.6 | 5.3 | 470 | 100.0 | 2.5 | 7 |
| Pos Culture | AutoTuner | 38.3 | 3.6 | 3.0 | 66.7 | 292.0 | 2.0 | 3 |
| Neg Standards | IPO | 26.0 | 12.0 | 6.2 | 250.0 | 100.0 | 3.6 | 10 |
| Neg Standards | AutoTuner | 29.3 | 5.7 | 4.0 | 436.8 | 1421.3 | 2.0 | 6 |
| Neg Culture | IPO | 60.0 | 27.4 | 4.7 | 121.0 | 100.0 | 4.0 | 9 |
| Neg Culture | AutoTuner | 66.9 | 4.9 | 1.8 | 7.8 | 117.5 | 2.0 | 3 |

**Table S3.** Feature count from each dataset during the different stages of quality assurance processing of culture data. The initial feature count was reduced after processing to remove blanks ('post blank)', features found in only one replicate ('post reproducibility), isotopologues and adducts ('post isotopes', and 'post adducts', respectively), and features with a CV greater than 0.4 in the pooled samples ('post CV').

| Ionization Mode | Algorithm | Initial Feature Count | Post Blank | Post Reproducibility | Post Isotopes | Post Adducts | Post CV |
|---|---|---|---|---|---|---|---|
| Negative | IPO | 40422 | 37903 | 8225 | 7695 | 4324 | 4226 |
| Negative | AutoTuner | 22599 | 17640 | 2921 | 2805 | 1444 | 1363 |
| Positive | IPO | 28794 | 28042 | 5907 | 5591 | 3628 | 3520 |
| Positive | AutoTuner | 13731 | 12451 | 2099 | 2012 | 1225 | 1143 |

**Table S4.** Counts of total detected features with MS/MS within figures 3 and S5 Venn diagrams.

| Ionization Mode | AutoTuner $MS^2$ Count | IPO $MS^2$ Count | Intersect $MS^2$ Count |
|---|---|---|---|
| Positive | 122 | 686 | 477 |
| Negative | 115 | 448 | 197 |

**Table S5.** Standard parameters used within centWave algorithm and their number of possible combinations. We cite these values in our discussion of speed improvements gained via AutoTuner relative to traditional parameter sweeping approaches dependent on optimization functions.

| **Parameter** | **Type** | **Possible Choices** | **Checked by AutoTuner** |
|---|---|---|---|
| ***ppm*** | Continuous | Infinite | Yes |
| ***S/N Threshold*** | Continuous | Infinite | Yes |
| ***Scan count*** | Continuous | Infinite | Yes |
| ***Noise*** | Continuous | Infinite | Yes |
| ***Prefilter intensity*** | Continuous | Infinite | Yes |
| ***Minimum Peak-width*** | Continuous | Infinite | Yes |
| ***Maximum Peak-width*** | Continuous | Infinite | Yes |
| ***mzDiff*** | Continuous | Infinite | No |
| ***Fit gauss*** | Boolean | 2 | No |
| ***Mz center function*** | Discrete | 4 | No |
| ***Integrate*** | Discrete | 2 | No |

**Table S6.** Number of unique features observed after processing data with unique mzDiff values. Columns two and three denote the mzDiff values used during pairwise comparisons of feature tables. Missing Count column represents the number of features observed outside the intersect of both feature tables. Feature tables were generated from 8 negative ion mode community data samples.

| Missing Count | mzDiff value of First Feature Table | mzDiff value of Second Feature Table |
|---|---|---|
| 0 | -0.001 | -0.002 |
| 0 | -0.002 | -0.003 |
| 0 | -0.003 | -0.004 |
| 0 | -0.004 | -0.005 |
| 0 | -0.005 | -0.006 |
| 0 | -0.006 | -0.007 |
| 0 | -0.007 | -0.008 |