Supporting Information

High-Resolution Filtering for Improved Small Molecule Identification via GC/MS

Nicholas W. Kwiecien[†‡], Derek J. Bailey[†‡], Matthew J.P. Rush[†‡], Jason S. Cole[§], Arne Ulbrich[†‡], Alexander S. Hebert[†], Michael S. Westphall[†], Joshua J. Coon[†‡]*

† Genome Center of Wisconsin, Madison, Wisconsin 53706, United States

‡ Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin, 53706, United States

§ Thermo Fisher Scientific, Austin, Texas 78728, United States

*Corresponding Author: jcoon@chem.wisc.edu

SUPPORTING INFORMATION

Urine Drug Analysis

The following GC gradient was used: 2.5 min isothermal at 60 ºC, ramp to 210 ºC at 40 ºC/min, ramp to 267 ºC at 5 ºC/min, ramp to 310 ºC at 40 ºC/min, then 6.2 min isothermal at 310 ºC. The MS transfer line and source temperatures were held at 280 ºC and 200 ºC, respectively. The mass range from 50-500 *m/z* was mass analyzed using a resolution of 30,000 ($m/\Delta m$), relative to 200 *m/z*. The AGC target was set to 1e6, and electron ionization (70 eV) was used. Lock mass calibration was employed during acquisition of these data. An unanticipated error occurred in calculation of the necessary mass correction, and many scans acquired during these experiments resulted in extreme mass errors (~25ppm). Large distortions in mass accuracy largely inhibit the described HRF approach. As such, during data processing each spectrum was restored to its native-state by removing the applied mass correction as reported in each scan header. Subsequent analyses did not employ this lock-mass correction and mass accuracy was unaffected.

Preparation of a *Saccharomyces cerevisiae* metabolite extract

Saccharomyces cerevisiae was grown on media containing dextrose and glycerol. $1x10^8$ cells were isolated by rapid vacuum filtration with a nylon filter membrane, washed with phosphate buffered saline, and submerged into a precooled 1.5 mL plastic tube containing a 2:2:1 acetonitrile/methanol/H_2O mixture.

Pesticide Analysis

The mixture containing 37 EPA 525.2 pesticides was diluted from 500 µg/mL to a working concentration of 3 ng/µL in acetone. A 1 µL aliquot was injected using a 1:10 split at a temperature of 275 ºC and separated at 1.2 mL/min He. The following GC oven gradient was used: isothermal at 100 ºC for 1 min, 8 ºC/min to 320 ºC, and isothermal at 320 ºC for 3 min. Transfer line and

source temperatures were maintained at 275 ºC and 225 ºC, respectively. In each MS scan, the range from 50-650 *m/z* was analyzed using a resolution of 17,500 (*m/*Δ*m*), relative to 200 *m/z*. Maximum injection times of 100 ms were allowed at an AGC target of 1e6. Electron ionization (EI) at 70 eV was used.

Additional Reference Standard Analysis

Stock solutions for all other reported standards were prepared individually at a concentration of 1 mg/mL in appropriate solvents. Mixtures containing ~5-10 reference standards were prepared by combining 20 µL aliquots of each standard using no specific organizational scheme. These mixtures were dried down under nitrogen, resuspended in 100 µL of the MSTFA + 1% TMCS derivatization reagent, capped, vortexed, and heated at 60 ºC for 15 minutes. 100 µL of ethyl acetate was then added to each mixture before being transferred to an autosampler vial. The same GC oven gradient and MS parameters as described in *Urine Drug Analysis* were also used here.

Spectral Deconvolution

Following data collection raw EI-MS spectral data was deconvolved into 'features' and then grouped into individual spectra containing only product ions stemming from a singular parent. This step was critical as the inclusion of extraneous fragment ions in a spectrum can diminish the ability of the algorithm to annotate all observed peaks with exact chemical formulas constrained by the atom set of the parent. Every peak in the raw data file was considered. Peaks observed in at least five consecutive scans having *m/z* values within ±10 ppm of their averaged *m/z* were grouped together as a data feature. Note that mass accuracy is a function of and S/N, and ppm tolerance a function of m/z. The 10 ppm tolerance was empirically observed to yield complete chromatographic profiles which were free of interference from neighboring peaks. Peaks were added successively to these groups and the average *m/z* value was recalculated after each

addition. Following aggregation of peaks into features, smoothed intensity profiles were created for each. Spurious features arising from noise were eliminated from consideration by requiring that each feature exhibit a “peak-like” shape. All features were required to rise to an apex having at least twice the intensity of the first and last peaks included. Any features arising from fragments common to closely eluting precursors were split into separate features at significant local minima. Features reaching an elution apex at approximately the same time were grouped together. Features were first sorted based on apex intensity. Starting with the most intense fragment a discrete time window around the apex was created. All features having an apex within this window were then grouped together. The width of this window was set to include all peaks having an intensity ≥ 96% of the apex peak’s intensity as a default. More conservative criteria was used for the extraction of spectra in the urine drug spike-in and discovery metabolomics experiments given the complex background. Here the time window was set to include peaks having an intensity ≥ 99% of the apex. Following feature grouping, a new spectrum was created for each group and populated with peaks representing each feature in the group. Peak *m/z* and intensity values were set equal to the intensity-weighted *m/z* average of all peaks in the corresponding feature and the intensity at the apex, respectively.

Small Molecule Identification via Spectral Matching

Compound identifications for the small molecules analyzed were assigned by comparing deconvolved high-resolution spectra against unit-resolution reference spectra present in the NIST 12 MS/EI Library. All 212,961 unit-resolution reference spectra in the library were exported to a .JDX file through the NIST MS Search 2.0 program and converted to a format suitable for matching against acquired Q Exactive GC spectra. A pseudo-unit resolution copy of each high-resolution spectrum was created by combining the intensities of peaks falling within the same nominal mass range. The nominal mass value was reported as peak *m/z* and all intensity values were normalized relative to the spectrum’s base peak (set to 999). To calculate spectral similarity

between experimental and reference spectra a weighted dot product calculation was used. First, all peaks in a spectrum were scaled using the following normalization factors reported in the literature which were determined to provide optimal spectral matching results[1]:

$$m/z_{normalized} = m/z_{measured} \times 1.3$$

$$intensity_{normalized} = intensity_{measured}^{0.53}$$

These normalization factors redistribute the weight placed on any given spectral peak in two ways: First, by scaling m/z by a factor of 1.3x, more massive peaks (which are inherently more diagnostic for spectral matching) are given greater weight. Second, by scaling intensity by a factor of $x^{0.53}$ more intense peaks are given relatively less weight. This is done to ensure that no single peak can disproportionately influence spectral matches. The described normalizations were applied to all reference spectra as well. The following dot product equation was used to measure spectral similarity:

$$100 \times \frac{\sum(m/z[Intensity_{experimental} * Intensity_{reference}]^{0.5})^2}{\sum(Intensity_{experimental} * m/z)\sum(Intensity_{reference} * m/z)}$$

Although simplistic, this approach was more than adequate for retrieving candidate compounds having similar fragmentation patterns to experimentally derived spectra. To increase search space as much as possible all reference spectra were matched against each unit resolution copy of a Q Exactive GC spectrum in the 'discovery metabolomics analysis'. All compounds reported yielded a confident spectral match with a reference spectrum in the NIST database.

High-Resolution Filtering: Theoretical Fragment Generation

A set of theoretical fragments for each candidate compound was produced by generating all unique combinations of atoms from the set contained in the parent chemical formula which can be calculated by:

$$x = \sum_{i}^{n} (i_a + 1)$$

where *x* is the number of theoretical fragments stemming from a given chemical formula, n is the number of unique elements in the formula, and i_a represents the atom count of that element within the formula. The most abundant isotope for each atom was used with the exception of bromine and chlorine. ^{79}Br and ^{81}Br have natural isotopic abundances of 0.5069 and 0.4931, respectively. Similarly, ^{35}Cl and ^{37}Cl have natural abundances of 0.7576 and 0.2424. For each theoretical fragment containing either a bromine or chlorine an additional variant was generated where a heavier isotope was exchanged for its lighter counterpart. This process was repeated in a combinatorial manner for those theoretical fragments containing multiple Br and/or Cl atoms. Generation of additional isotopic theoretical fragments for those candidates containing atoms in the set {^{12}C, ^{32}S, ^{28}Si} was done on a case-by-case basis during the theoretical fragment/peak matching process.

High-Resolution Filtering: Theoretical Fragment/Peak Matching

It is assumed that all fragment peaks in an EI-MS spectrum are radical cations. Accordingly, the mass of an electron was subtracted from the monoisotopic mass of each fragment in the set of candidates. Starting with the least massive peak in the Q Exactive GC spectrum, theoretical fragments falling within a ± 10 ppm tolerance centered around the peak's measured *m/z* were found. This tolerance was empirically determined to be the optimal allowed mass tolerance as it enabled annotation of low S/N fragments where mass accuracy is diminished while maintaining discrimination against spurious chemical formulas (**Supplementary Figure 6**). If no fragments were present within this range, the algorithm moved to the next most massive peak and repeated the process. If a single fragment was found within this range, isotopic variants containing substituted ^{13}C, ^{33}S, ^{34}S, ^{29}Si, or ^{30}Si atoms were generated where appropriate and added to the list of candidate fragments. If multiple fragments were found within the allowed tolerance each

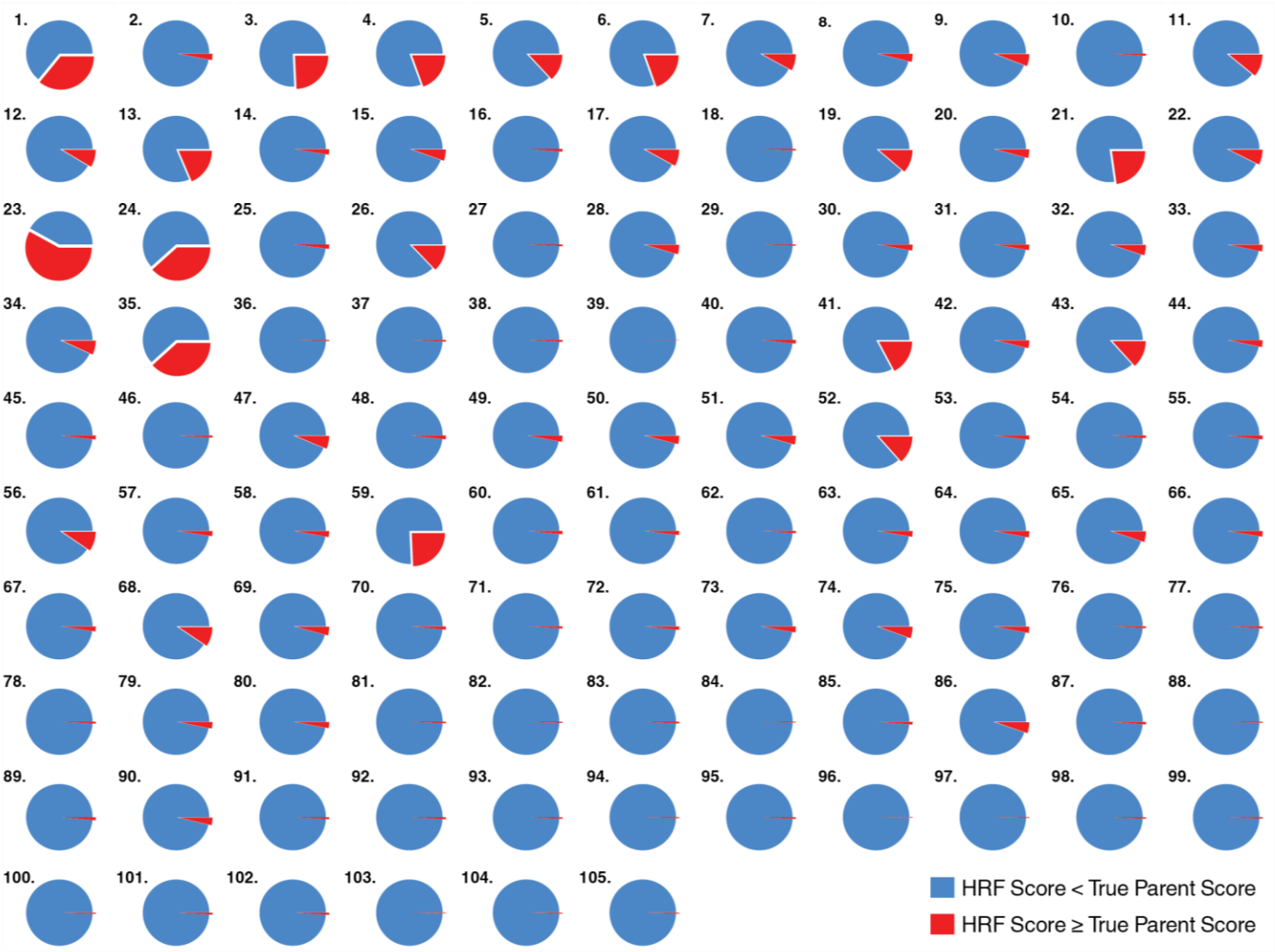
fragment was independently evaluated to determine how many additional peaks/signal could be matched. The theoretical fragment resulting in the largest amount of additional matched signal was assumed to be correct and substituted isotopic theoretical fragments were added to the list of candidate theoretical fragments. All peaks which had matching theoretical fragments were stored. After all peaks were considered the total ion current that was matched to a theoretical fragment as calculated by:

$$\sum(mz * intensity)_{annotated} / \sum(mz * intensity)_{observed}$$

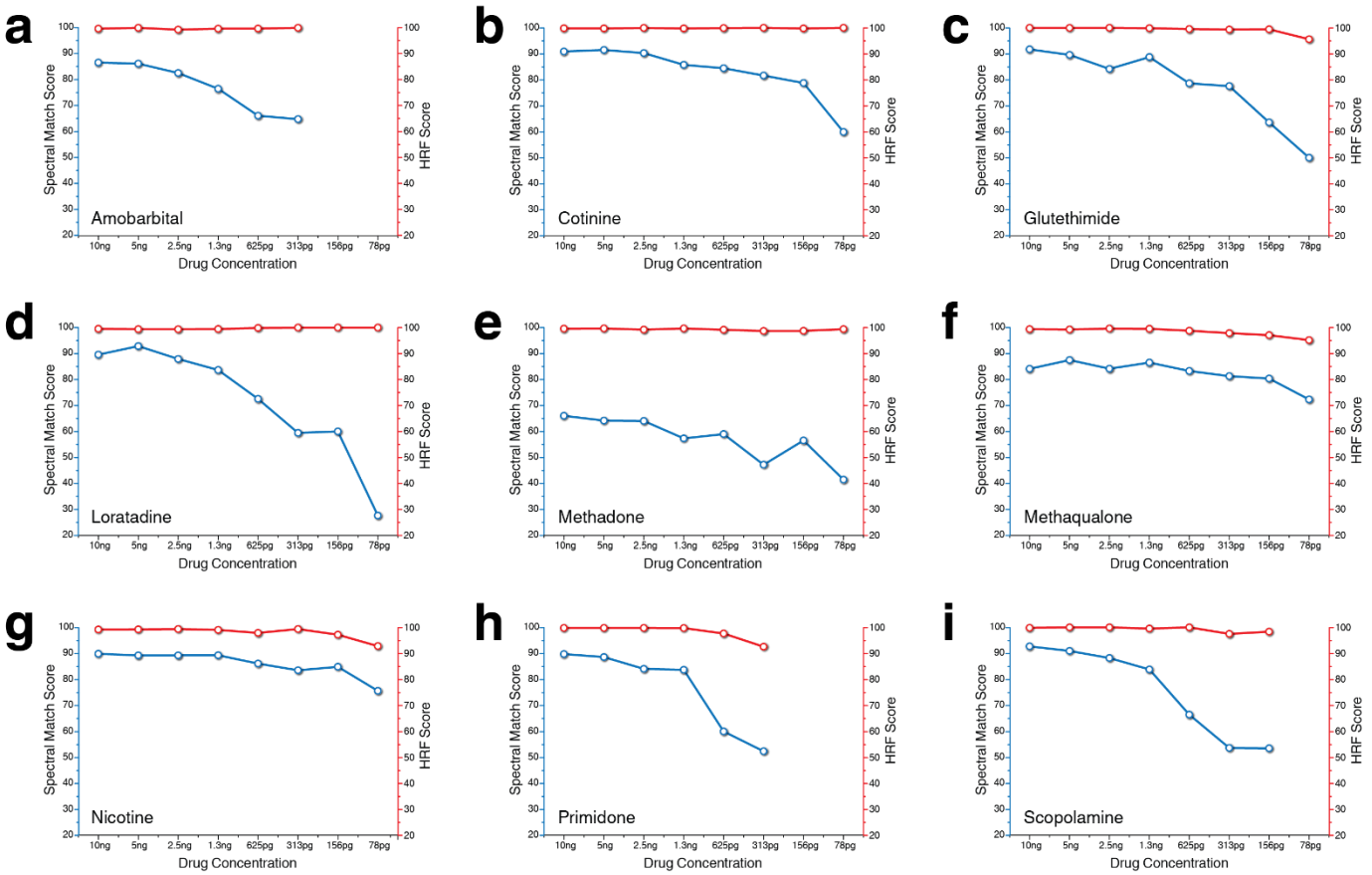
was returned. This scoring calculation was deemed appropriate as it gives additional weight to larger ions which are inherently more diagnostic of a given precursor than less massive ions. Conceptually, there are fewer molecules in existence which can theoretically produce a fragment at 300 *m/z* than there are which can produce a fragment at 200 *m/z*. An analysis of execution time (on a desktop PC) of the high-resolution filtering process using 232 metabolite spectra and 50 candidate matches to each spectrum is highlighted in **Supplementary Figure 7**.

References

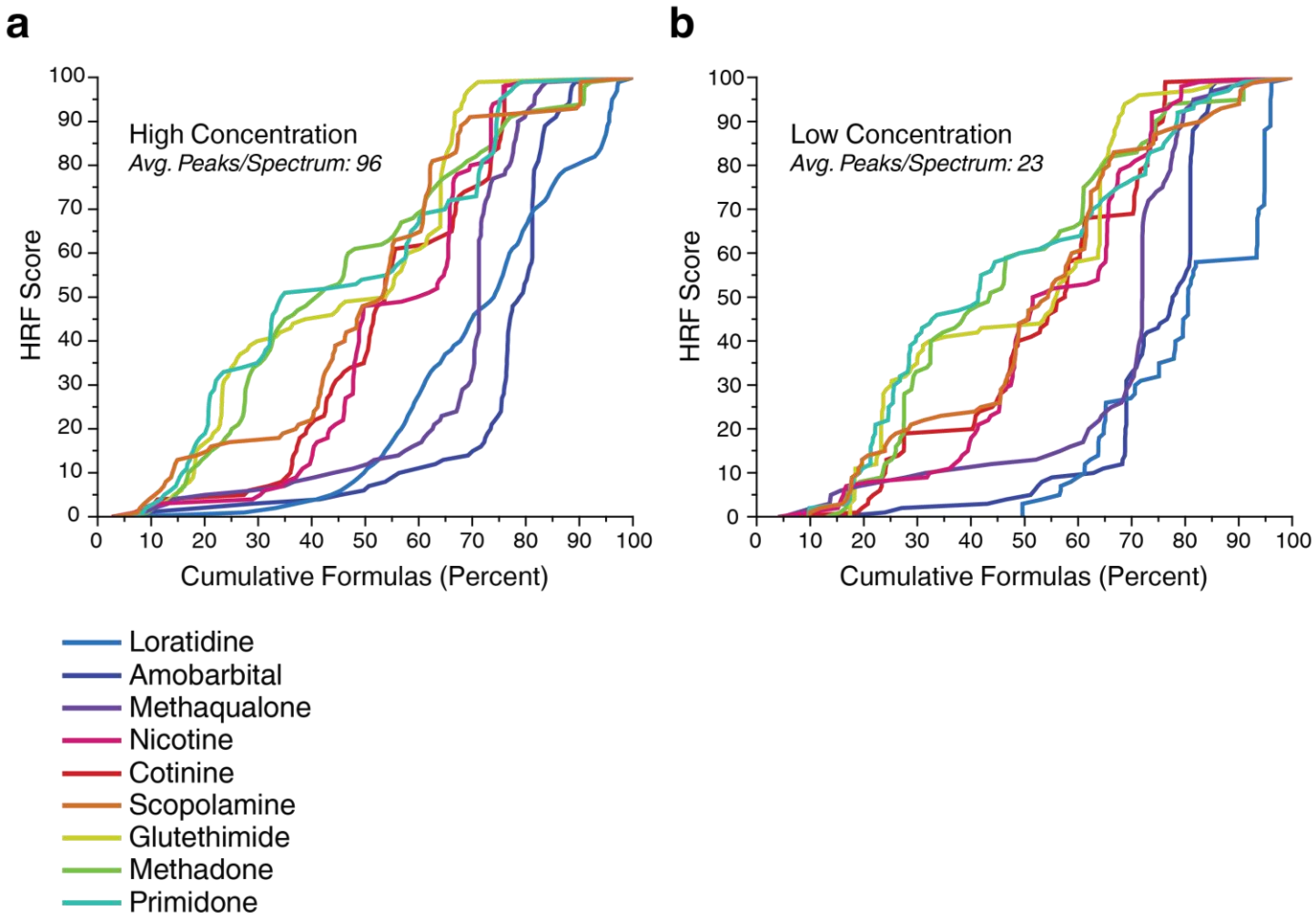
(1) Kim, S.; Koo, I.; Wei, X.; Zhang, X. *Bioinformatics* **2012**, *28* (8), 1158–1163.



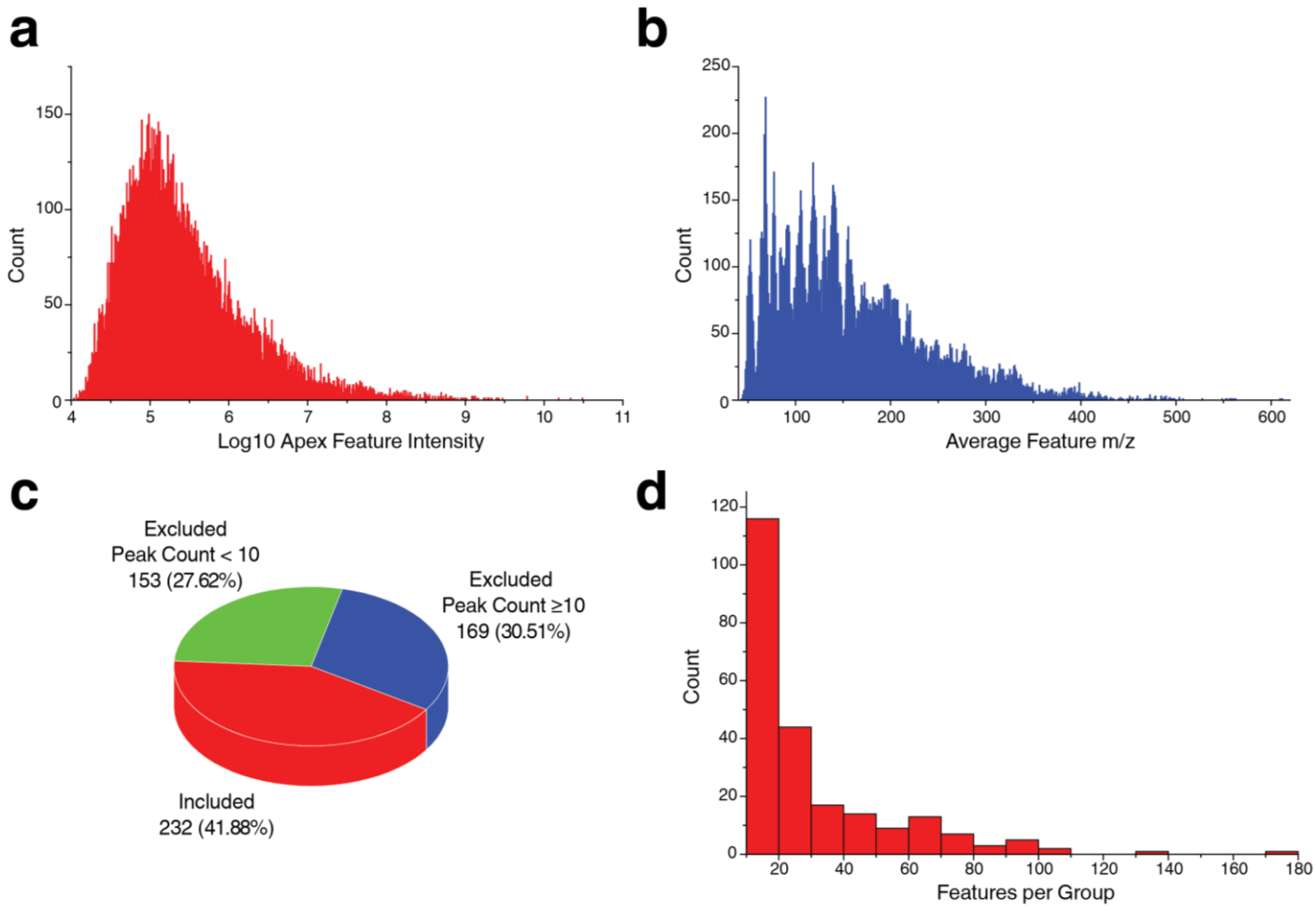
Supplementary Figure 1. Global high-resolution filtering results. For all 105 reference spectra analyzed in this study 60,560 HRF scores were calculated using a unique chemical formulas from the NIST 12 EI reference library. Shown here are the results of that analysis for all reference spectra (1-105) ordered by increasing monoisotopic mass. The calculated scores are separated into two categories; formulas yielding HRF scores less than the true parent score (blue), and formulas yielding HRF scores greater than or equal to the true parent score (red). More detailed results are shown in **Supplementary Table 2**. Note that for the majority of considered spectra a very small percentage of formulas can produce a similarly high (or higher score) with few exceptions. Cursory analysis of the cases where a large percentage of formulas can produce high-quality results (1, 23, 24, 35.) indicates that such compounds tend to have more simplistic formulas ($C_{10}H_{15}N$, $C_{12}H_{14}N_2O_2$, $C_{15}H_{10}O_2$, $C_{16}H_{17}NO$, respectively). We note that these compounds are comprised exclusively of the four most common organic elements, namely carbon, hydrogen, nitrogen, and oxygen. For compounds with increased chemical complexity the method exhibits increased specificity, as anticipated.



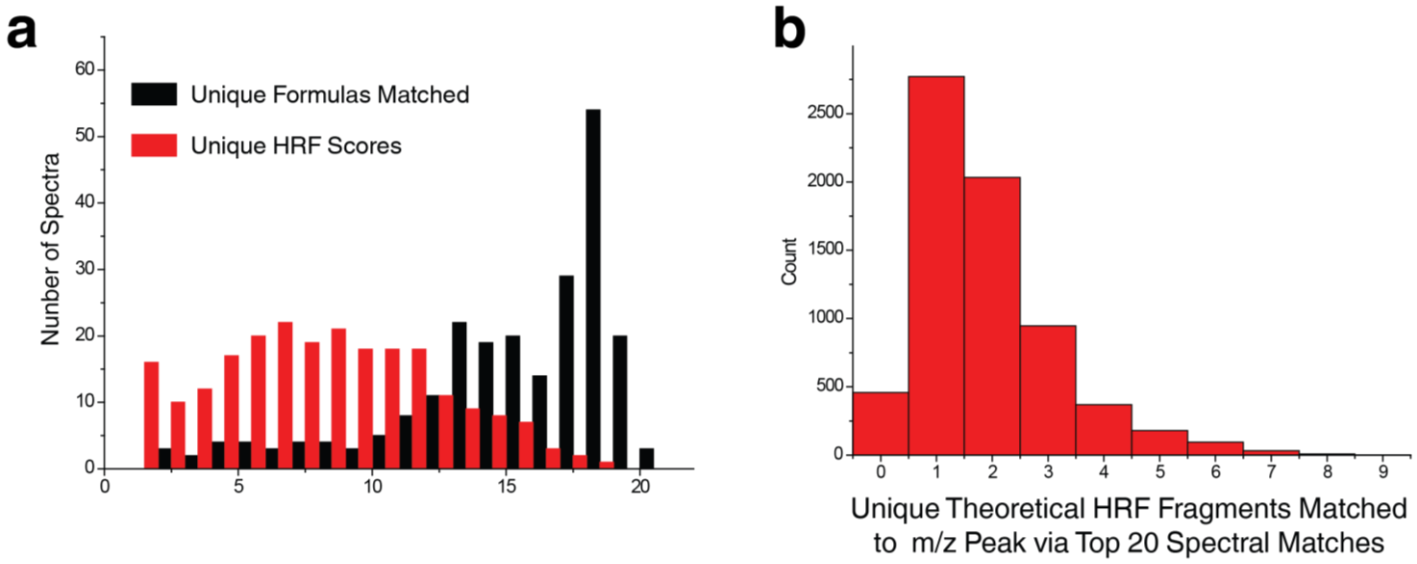
Supplementary Figure 2. Individual analyses of drugs spiked into human urine at variable concentration. (**a-i**) Shown here are the measured spectral match and HRF scores for all deconvolved spectra extracted from the urine spike-in data set. These data are the same as that shown in **Fig. 3b**. Corresponding spectral match and HRF score lines are plotted together for clarity. It is noted that at reduced concentrations observed spectral match score tends to decline while the HRF metric remains high.



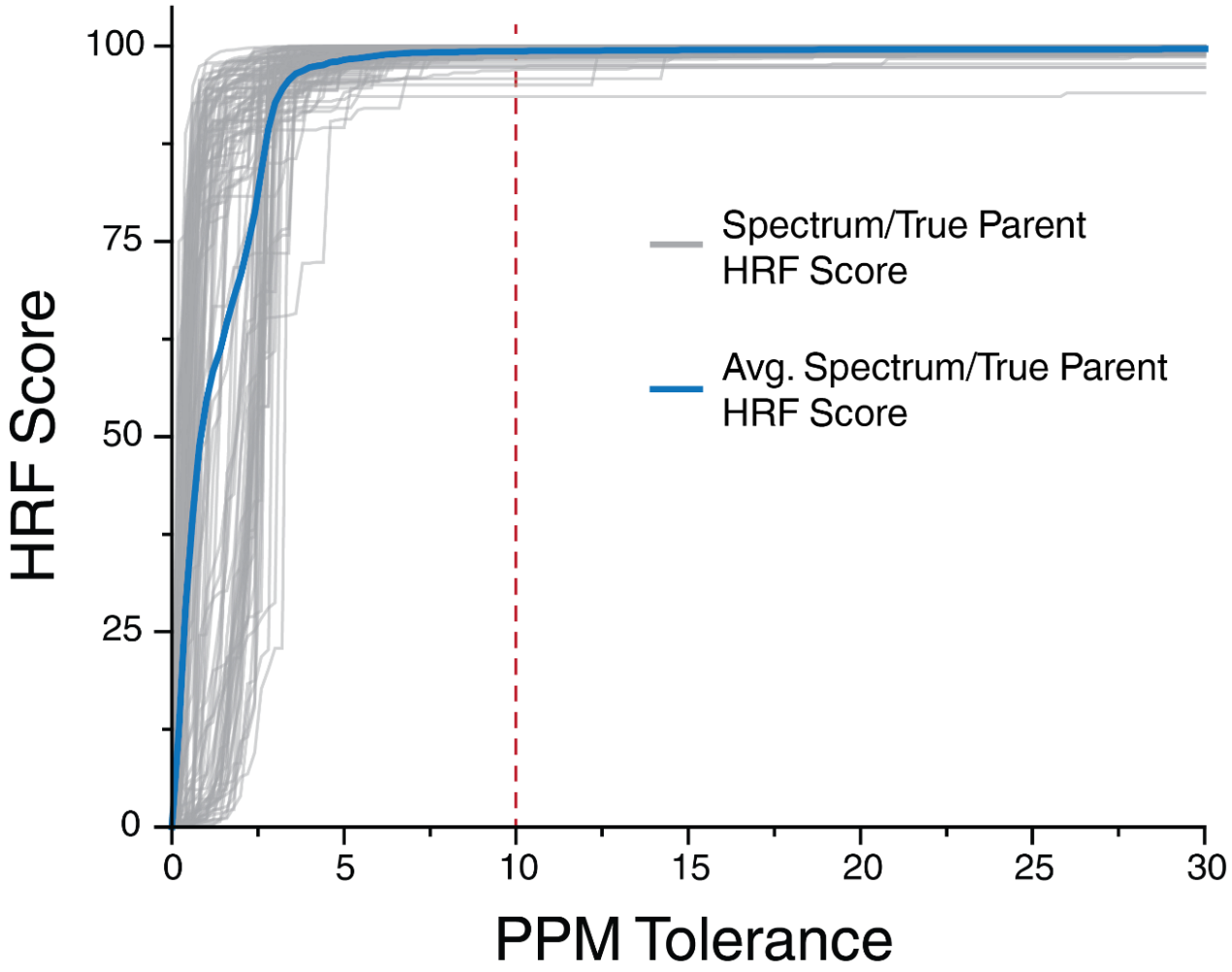
Supplementary Figure 3. HRF Specificity. Two spectra for each of the drugs analyzed were extracted, one at the highest measured concentration and one at the lowest. Given that these drugs are relatively small these formulas were assumed to more accurately reflect a pool of potential candidate molecules, rather than utilizing all formulas in the database. 55,229 HRF scores were calculated using unique formulas (0-500 Da) from the NIST 12 EI reference library. Cumulative distributions of these scores are shown for each spectrum at high concentration (**a**) and low concentration (**b**). These data are the same as that shown in **Figure 3d** but are color-coded here for clarity. The specificity of the method does not appear to change whether a "peak-rich" or a "peak-depleted" spectrum is considered as similar cumulative curves are generated for each drug. This data suggests that even spectra collected at diminished concentrations will contain sufficient information for the method to maintain specificity.



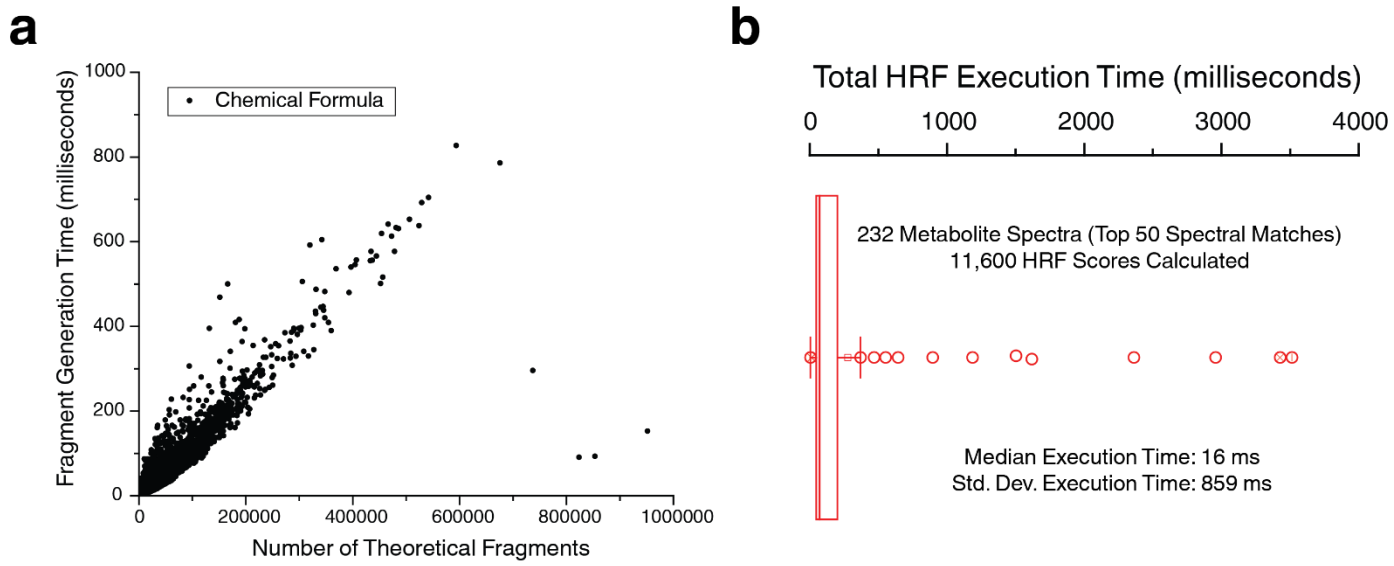
Supplementary Figure 4. Discovery Metabolomics Dataset Overview. Deconvolution of a raw data file from the 30-minute analysis of an extracted/TMS-derivatized yeast metabolome yielded 19,367 features which met the requirements for consideration as a true analyte feature. The distribution of feature intensities and m/z values are shown above (**a, b**). These extracted features were subsequently placed into 554 groups. For our analyses we isolated only those feature groups which contained 10+ peaks and were not found in a corresponding background run. The distribution of included/excluded feature groups in shown in **c**. The 232 feature groups (read: spectra) included in our analyses were assumed to be biological in nature and contained a median of 20 features per group (**d**).



Supplementary Figure 5. HRF Specificity in Discovery Metabolomic Analysis. For each of the 232 metabolite spectra in our dataset the top 20 spectral matches were retrieved using a database search, and a corresponding HRF score was calculated for each. The uniqueness of these 20 matches with regards to chemical formula and associated HRF score are shown in **a**. Given these distributions, it is apparent that many formulas which are chemically inequivalent can produce identical HRF scores. We predicted that in such instances, individual peaks were being annotated with conserved subsets of atoms from different formula precursors. For each *m/z* peak in each spectrum considered, we show the distribution of unique annotations assigned to that peak from all 20 matched precursors (**b**). These data show that often only a single formula annotation is ever assigned to a given *m/z* peak suggesting that only formulas containing the appropriate set of atoms from a given precursor will be able to achieve a high score.



Supplementary Figure 6. HRF Theoretical-Fragment-to-Peak Matching Mass Tolerances. Using the set of 105 spectra from pure reference standards we calculated HRF scores from the true parent chemical formula using allowed mass tolerances ranging from 0-30 ppm. The gray curves above highlight the associated score at a given ppm tolerance for each spectrum. The curve in blue is the average of all 105 curves at each data point. Ideally this tolerance is kept very small as to prevent spurious annotations from being assigned. However PPM tolerance width is a function of *m/z* and we acknowledge that mass accuracy is diminished in times of reduced S/N. Based on these data we opted to use a 10 ppm mass tolerance for all analyses.



Supplementary Figure 7. HRF Execution Time. To demonstrate the feasibility of the HRF approach for routine discovery metabolomic data analysis we characterize the total time needed to generate all theoretical fragments from 60,560 different chemical formula inputs (a). We find a linear relationship between fragment generation time and the number of theoretical fragments and note that nearly 1e6 theoretical fragments can be generated in less than one second. Additionally, we characterize the total HRF execution time (theoretical fragment generation + theoretical fragment-peak matching) using the top 50 matched formulas to 232 metabolite spectra (11,600 HRF scores in total) in **b**. The box designates the innerquartile range (IQR) and the whiskers represent 1.5x the upper/lower IQR, respectively. Open-circles represent outliers. Here we find a median total HRF execution time of 16 ms with a standard deviation of 859 ms. All analyses described in this work were carried out on a personal computer with an Intel I5-4570 3.2 GHz quad-core processor and 16 GB of RAM running Windows 7 Professional.

Supplementary Table 1. Shown here are results from all analyzed reference compounds complete with raw file name, retention time, HRF score, spectral match score, peak count, and the reference spectrum name as reported in NIST 12.

Name	HRF Score	Spectral Match Score	Peak Count	Proper Name (NIST 12 EI Database)
2'-Deoxyadenosine	100	80.23787	121	2'-Deoxyadenosine, N-trimethylsilyl-, bis(trimethylsilyl) ether
6-Aminocaproic Acid	99.85167	73.04963	114	Hexanoic acid, 6-amino-, bis(trimethylsilyl) deriv,
Acetaminophen	98.99406	85.06104	115	Acetamide, N-(trimethylsilyl)-N-[4-[(trimethylsilyl)oxy]phenyl]-
Adenine	98.48893	88.66699	90	9H-Purin-6-amine, N,9-bis(trimethylsilyl)-
Adenosine	100	81.29393	117	Adenosine-tetrakis(trimethylsilyl)-
Alachlor	100	78.14022	124	Alachlor
Alanine	98.73187	84.82428	42	l-Alanine, trimethylsilyl ester
Ametryn	99.37576	83.82522	125	Ametryn
Amobarbital	97.61185	86.09109	91	Amobarbital
Ascorbic Acid	99.95632	81.42812	162	L-Ascorbic acid, 2,3,5,6-tetrakis-O-(trimethylsilyl)-
Aspartic Acid	100	87.35514	84	L-Aspartic acid, N-(trimethylsilyl)-, bis(trimethylsilyl) ester
Atraton	99.50053	85.15589	110	Atraton
Atrazine	99.71586	86.05622	108	Atrazine
Beta-Alanine	98.84262	73.69351	52	,beta,-Alanine, N-(trimethylsilyl)-, trimethylsilyl ester
Beta-Sitosterol	99.92321	85.28424	184	,beta,-Sitosterol trimethylsilyl ether
Bromacil	99.84644	84.28455	70	Bromacil
Butachlor	99.91863	80.29282	115	Butachlor
Butylate	98.88798	65.56806	60	Carbamothioic acid, bis(2-methylpropyl)-, S-ethyl ester
Caffeine	99.61229	85.29047	88	Caffeine
Catechin	99.92232	62.57484	111	2H-1-Benzopyran, 3,4-dihydro-2-[3,4-bis[(trimethylsilyl)oxy]phenyl]-3,5,7-tris[(trimethylsilyl)oxy]-, (2R-trans)-
Chlorpropham	99.96756	88.86683	61	Chlorpropham
Cotinine	99.74813	90.64544	105	Cotinine
Cyanazine	99.91903	82.52818	134	Cyanazine
Cycloate	99.07497	75.41157	68	Cycloate
Cysteine	99.9446	86.59517	54	L-Cysteine, N,S-bis(trimethylsilyl)-, trimethylsilyl ester
Cystine	100	82.68418	76	L-Cystine, N,N'-bis(trimethylsilyl)-, bis(trimethylsilyl) ester
Diphenamid	95.06315	73.17383	48	Diphenamid
Diphenhydramine	99.86228	76.05572	51	Acetamide, 2,2-diphenyl-N-(2-dimethylamino)ethyl-

Dopamine	99.68245	86.51747	119	Silanamine, N-[2-[3,4-bis[(trimethylsilyl)oxy]phenyl]ethyl]-1,1,1-trimethyl-
EPTC	98.66519	74.36759	44	Carbamothioic acid, dipropyl-, S-ethyl ester
Estriol	99.96204	69.27833	137	Tri(trimethylsilyl) derivative of estriol
Estrone	99.49286	84.59311	168	Trimethylsilylestrone
Etridiazole	100	86.52784	80	Etridiazole
Fenarimol	99.69995	78.49869	123	Fenarimol
Ferulic Acid	98.61093	82.55173	147	Trimethylsilyl 3-methoxy-4-(trimethylsilyloxy)cinnamate
Flavone	97.29626	89.69236	79	Flavone
Fluridone	97.01718	81.5551	123	Fluridone
Fumaric Acid	98.6845	53.11481	37	2-Butenedioic acid (Z)-, bis(trimethylsilyl) ester
Gamma Aminobutryic Acid	100	64.91472	14	Butanoic acid, 4-[(trimethylsilyl)amino]-, trimethylsilyl ester
Glucosamine	100	85.60832	141	Glucosamine per-TMS
Glucose	100	86.02583	98	Glucopyranose, 1,2,3,4,6-pentakis-O-(trimethylsilyl)-, D-
Glutamic Acid	99.58506	86.86825	96	Glutamic acid, N-(trimethylsilyl)-, bis(trimethylsilyl) ester, L-
Glutamine	100	78.12936	96	l-Glutamine, tris(trimethylsilyl) deriv,
Glutaric Acid	99.88249	65.13565	54	Pentanedioic acid, bis(trimethylsilyl) ester
Glutethimide	99.55617	92.58142	110	Glutethimide
Glyceric Acid	100	80.20763	81	Propanoic acid, 2,3-bis[(trimethylsilyl)oxy]-, trimethylsilyl ester
Glycine	100	72.05176	33	Glycine, N,N-bis(trimethylsilyl)-, trimethylsilyl ester
Hexazinone	99.46783	82.67615	72	1,3,5-Triazine-2,4(1H,3H)-dione, 3-cyclohexyl-6-(dimethylamino)-1-methyl-
Histidine	100	75.48915	63	L-Histidine, N,1-bis(trimethylsilyl)-, trimethylsilyl ester
Homovanillic Acid	99.54148	81.13459	81	Trimethylsilyl [3-methoxy-4-(trimethylsilyloxy)phenyl]acetate
Inositol	100	61.85832	135	Myo-Inositol, pentakis-O-(trimethylsilyl)-
Isoleucine	99.69393	86.31592	91	L-Isoleucine, N-(trimethylsilyl)-, trimethylsilyl ester
Ketamine	99.1702	91.45966	147	Ketamine
L (+) Lactic Acid	99.80252	73.85199	57	D-(-)-Lactic acid, trimethylsilyl ether, trimethylsilyl ester
L-2 Aminobutyric Acid	99.75521	85.93663	53	l-2-Aminobutyric acid, N-trimethylsilyl-, trimethylsilyl ester
Loratidine	99.26171	89.68975	153	Loratadine
Lysine	100	52.51087	90	L-Lysine, N2,N6,N6-tris(trimethylsilyl)-, trimethylsilyl ester
Mandelic Acid	99.69772	91.22946	66	Benzeneacetic acid, ,alpha,-[(trimethylsilyl)oxy]-, trimethylsilyl ester
Mescaline	99.78119	91.25275	77	Acetamide, N-(3,4,5-trimethoxyphenethyl)-

Metaqualone	98.63943	88.19924	129	Methaqualone
Methadone	99.18112	64.81793	115	Methadone
Methamphetamine	98.85648	66.2167	27	Methamphetamine
Methylmalonic Acid	99.76899	61.44021	38	Propanedioic acid, methyl-, bis(trimethylsilyl) ester
Metolachlor	100	87.14172	72	Metolachlor
Metribuzin	95.83894	78.23404	126	Metribuzin
MGK-264	100	67.25826	95	N-(2-Ethylhexyl)-5-norbornene-2,3-dicarboximide
Minoxidil	99.86569	94.87978	118	Desoxy-minoxidyl
Molinate	98.57083	77.33713	48	Molinate
Napropamide	98.81199	80.58035	72	Napropamide
Naproxen	99.14971	88.82363	69	2-Naphthaleneacetic acid, 6-methoxy-,alpha,-methyl-, trimethylsilyl ester, (+)-
Nicotine	99.30713	90.8779	103	Pyridine, 3-(1-methyl-2-pyrrolidinyl)-, (S)-
Norflurazon	99.73092	83.5459	109	Norflurazon
Ornithine	99.63999	80.92918	142	Ornithine, tri-TMS
Orotic Acid	100	42.59934	33	4-Pyrimidinecarboxylic acid, 2,6-bis(trimethylsiloxy)-, trimethylsilyl ester
Oxalic Acid	98.7125	65.73171	30	Ethanedioic acid, bis(trimethylsilyl) ester
Pebulate	97.36806	74.74838	56	Pebulate
Pipecolinic Acid	99.5349	81.8888	75	2-Piperidinecarboxylic acid, 1-(trimethylsilyl)-, trimethylsilyl ester
Primidone	99.88732	92.33499	95	Primidone
Proline	99.53685	67.4245	64	L-Proline, 1-(trimethylsilyl)-, trimethylsilyl ester
Prometon	99.46725	83.18783	76	Prometon
Prometryn	99.02092	85.43111	113	Prometryn
Propachlor	99.42461	80.98082	65	Acetamide, 2-chloro-N-(1-methylethyl)-N-phenyl-
Propazine	99.65145	82.094	99	Propazine
Propyzamide	99.64317	78.40575	77	Propyzamide
Pyroxidine	100	86.25164	122	Pyridine, 2-methyl-3-(trimethylsilyloxy)-4,5-bis-[(trimethylsilyloxy)methyl]-
Sarcosine	99.01318	75.64516	57	Bis(trimethylsilyl)sarcosine
Serine	100	86.97745	83	Serine, N,O-bis(trimethylsilyl)-, trimethylsilyl ester
Simazine	100	77.02246	58	Simazine
Simetryn	99.65115	85.2555	130	Simetryn
Sinapic Acid	99.20565	67.30941	24	Cinnamic acid, 3,5-dimethoxy-4-(trimethylsiloxy)-, trimethylsilyl ester
Succinic Acid	98.34062	69.62375	87	Butanedioic acid, bis(trimethylsilyl) ester
Tebuthiuron	100	79.94081	58	Tebuthiuron
Terbacil	100	83.72495	47	Terbacil
Terbutryn	99.40774	84.2506	132	Terbutryn
Threonine	100	90.16955	122	N,O,O-Tris(trimethylsilyl)-L-threonine

trans-4-hydroxyproline	100	90.00911	78	L-Proline, 1-(trimethylsilyl)-4-[(trimethylsilyl)oxy]-, trimethylsilyl ester, trans-
Triadimefon	99.95845	69.92398	84	Triadimefon
Tricyclazole	93.4973	79.30223	63	Tricyclazole
Trifluralin	100	66.04019	196	Trifluralin
Tryptamine	98.85996	80.35281	108	1H-Indole-3-ethanamine, N,1-bis(trimethylsilyl)-
Tryptophan	99.9878	90.48896	72	L-Tryptophan, N,1-bis(trimethylsilyl)-, trimethylsilyl ester
Tyrosine	100	84.23964	97	L-Tyrosine, N,O-bis(trimethylsilyl)-, trimethylsilyl ester
Uridine	99.99264	74.19771	121	Uridine, tetra(trimethylsilyl)-
Valine	99.71247	89.14675	84	L-Valine, N-(trimethylsilyl)-, trimethylsilyl ester
Vernolate	98.48952	75.4259	56	Carbamothioic acid, dipropyl-, S-propyl ester

Supplementary Table 2. Global HRF analysis. Shown here is a summary of the returned HRF results when calculating scores for the 105 dataset spectra against 60,560 unique chemical formulas. Compounds are ranked by ascending monoisotopic mass. The raw number of formulas which produce a HRF score less than, or greater than or equal to the true parent are shown in columns labeled HRF < Parent Score and HRF >= Parent Score. Using the pool of formulas which yielded a HRF Score>= the true parent HRF score the number of true and false supersets were determined. A superset is a formula where all of the atoms in the true parent set are also contained. Non-supersets were those formulas which failed to meet this condition. For those non-supersets the average percentage of atoms shared with the true parent was calculated, along with the average and median number of additional atoms held by the formula in question. We find that these non-supersets which can achieve similarly high HRF scores as the true parent often share a large percentage of atoms with the correct precursor (93.574%) and contain a substantial number of additional atoms on average (19.506)

ID Number	Name	Chemical Formula	Monoisotopic Mass	HRF < Parent Score	HRF ? Parent Score	True Supersets	False Supersets	Percent of Atoms Shared (False Supersets)	Avg. Additional Atoms (False Supersets)	Median Additional Atoms (False Supersets)
1	Methamphetamine	C10H15N	149.1204	38804	21756	20004	1752	95.7785	11.5228	11
2	Alanine (TMS)	C6H15NO2Si	161.0872	58714	1846	1705	141	91.3475	17.6241	16
3	Nicotine	C10H14N2	162.1157	45856	14704	14081	623	95.9007	27.8042	25
4	Cotinine	C10H12N2O	176.095	48758	11802	10994	808	95.8515	23.3837	22
5	Molinate	C9H17NOS	187.1031	52685	7875	3271	4604	96.1847	29.7068	26
6	Tricyclazole	C9H7N3S	189.0361	48720	11840	3640	8200	92.2787	27.109	23
7	EPTC	C9H19NOS	189.1187	55743	4817	2610	2207	96.3883	27.836	24
8	Minoxidil	C9H15N5	193.1327	58223	2337	1272	1065	94.3694	29.3765	25
9	Caffeine	C8H10N4O2	194.0804	57003	3557	1999	1558	94.6834	28.1573	24
10	Simazine	C7H12ClN5	201.0781	59960	600	445	155	91.3548	29.0129	25
11	Pebulate	C10H21NOS	203.1344	53944	6616	2005	4611	93.5085	21.077	16
12	Vernolate	C10H21NOS	203.1344	55399	5161	2008	3153	93.3052	20.2851	14
13	Propachlor	C11H14ClNO	211.0764	49306	11254	2869	8385	95.9826	24.3171	21

14	Atraton	C9H17N5O	211.1433	58994	1566	1272	294	95.2594	28.6939	25
15	Chlorpropham	C10H12ClNO2	213.0557	57248	3312	2326	986	94.3634	17.3824	13
16	Simetryn	C8H15N5S	213.1048	59825	735	418	317	93.854	32.3849	29
17	Metribuzin	C8H14N4OS	214.0888	55724	4836	832	4004	91.6637	22.0844	18
18	Atrazine	C8H14ClN5	215.0938	60114	446	346	100	93.4643	25.81	23
19	Cycloate	C11H21NOS	215.1344	53755	6805	1966	4839	93.5488	19.554	14
20	Terbacil	C9H13ClN2O2	216.0666	58040	2520	1461	1059	91.5993	12.1681	10
21	Glutethimide	C13H15NO2	217.1103	46780	13780	11879	1901	95.1825	15.9495	13
22	Butylate	C11H23NOS	217.15	56103	4457	1534	2923	93.4305	19.6914	14
23	Primidone (TMS)	C12H14N2O2	218.1055	25420	35140	8596	26544	92.9682	22.3994	17
24	Flavone	C15H10O2	222.0681	37300	23260	19328	3932	92.4165	15.2411	13
25	Prometon	C10H19N5O	225.159	59327	1233	1022	211	95.2607	29.3507	26
26	Amobarbital	C11H18N2O3	226.1317	52802	7758	4579	3179	91.8019	12.2051	9
27	Ametryn	C9H17N5S	227.1205	60045	515	263	252	94.8413	31.0397	28
28	Tebuthiuron	C9H16N4OS	228.1045	57803	2757	674	2083	93.5979	14.1195	12
29	Propazine	C9H16ClN5	229.1094	60220	340	269	71	94.3662	27.3944	24
30	Beta-Alanine (TMS)	C9H23NO2Si2	233.1267	58845	1715	998	717	89.3211	18.7169	16
31	Sarcosine (TMS)	C9H23NO2Si2	233.1267	58980	1580	985	595	90.3747	19.5126	17
32	Oxalic Acid (TMS)	C8H18O4Si2	234.0744	57475	3085	1183	1902	90.2964	23.8312	19
33	Lactic Acid (TMS)	C9H22O3Si2	234.1107	58614	1946	1606	340	94.3301	20.4647	19
34	Ketamine	C13H16ClNO	237.092	56362	4198	2001	2197	96.5507	26.6359	22
35	Diphenamid	C16H17NO	239.131	37369	23191	11476	11715	90.584	13.4525	9
36	Cyanazine	C9H13ClN6	240.089	60253	307	167	140	92.734	26	22
37	Prometryn	C10H19N5S	241.1361	60093	467	235	232	95.1355	29.1853	26
38	Terbutryn	C10H19N5S	241.1361	60012	548	237	311	94.8002	26.9936	24
39	Etridiazole	C5H5Cl3N2OS	245.9188	60503	57	53	4	94.1176	27.5	29

40	L-2-Aminobutyric Acid (TMS)	C10H25NO2Si2	247.1424	59537	1023	807	216	93.7847	16.0463	14
41	Methaqualone	C16H14N2O	250.1106	50116	10444	8436	2008	94.7392	22.4158	17
42	Hexazinone	C12H20N4O2	252.1586	58238	2322	1556	766	96.2931	23.4021	20
43	Mescaline	C13H19NO4	253.1314	52518	8042	4640	3402	95.4717	21.1822	16
44	Propyzamide	C12H11Cl2NO	255.0218	58544	2016	1142	874	94.8216	21.7654	17
45	Proline (TMS)	C11H25NO2Si2	259.1424	59386	1174	893	281	93.9328	16.4484	15
46	Bromacil	C9H13BrN2O2	260.016	59918	642	493	149	91.9215	9.8121	9
47	Fumaric Acid (TMS)	C10H20O4Si2	260.09	56775	3785	1148	2637	89.227	21.1331	17
48	Valine (TMS)	C11H27NO2Si2	261.158	59442	1118	843	275	93.6406	14.8473	13
49	Methylmalonic Acid (TMS)	C10H22O4Si2	262.1057	58757	1803	1052	751	92.5258	25.1225	22
50	Succinic Acid (TMS)	C10H22O4Si2	262.1057	58114	2446	1110	1336	88.8946	21.1198	18
51	Alachlor	C14H20ClNO2	269.1183	57984	2576	730	1846	96.8609	24.0785	21
52	Napropamide	C17H21NO2	271.1572	52446	8114	6542	1572	95.3345	13.4135	11
53	Pipecolinic Acid (TMS)	C12H27NO2Si2	273.158	59364	1196	852	344	93.7962	15.8052	14
54	6-Aminocaproic Acid (TMS)	C12H29NO2Si2	275.1737	59818	742	594	148	94.2274	16.6081	14
55	Isoleucine (TMS)	C12H29NO2Si2	275.1737	59423	1137	795	342	93.3384	14.6316	13
56	MGK-264	C17H25NO2	275.1885	54814	5746	5135	611	96.1193	11.784	10
57	Glutaric Acid (TMS)	C11H24O4Si2	276.1213	59062	1498	1014	484	95.7821	22.6054	20
58	Adenine (TMS)	C11H21N5Si2	279.1335	58826	1734	69	1665	90.4166	27.5003	23
59	Diphenhydramine	C18H22N2O	282.1732	45835	14725	4299	10426	84.7088	7.9011	6
60	Metolachlor	C15H22ClNO2	283.1339	59613	947	514	433	95.888	11.7506	10
61	Glycine (TMS)	C11H29NO2Si3	291.1506	59405	1155	464	691	89.8855	18.4732	16

62	Triadimefon	C14H16ClN3O2	293.0931	59909	651	444	207	95.9608	20.6957	20
63	Acetaminophen (TMS)	C14H25NO2Si2	295.1424	58890	1670	856	814	93.0618	17.9853	16
64	Mandelic Acid (TMS)	C14H24O3Si2	296.1264	58718	1842	1294	548	93.2694	14.8467	12
65	Naproxen (TMS)	C17H22O3Si	302.1338	57397	3163	1658	1505	95.4431	18.5907	16
66	Norflurazon	C12H9ClF3N3O	303.0386	58917	1643	142	1501	92.7382	20.948	18
67	Tryptamine (TMS)	C16H28N2Si2	304.1791	59131	1429	389	1040	93.6819	19.0288	15
68	Methadone	C21H27NO	309.2093	54863	5697	3917	1780	95.1674	10.2607	9
69	Butachlor	C17H26ClNO2	311.1652	58015	2545	310	2235	97.1612	23.7154	20
70	Gamma Aminobutyric Acid (TMS)	C13H33NO2Si3	319.1819	59603	957	420	537	90.689	15.5512	14
71	Serine (TMS)	C12H31NO3Si3	321.1612	59945	615	337	278	93.5396	16.4209	14
72	Glyceric Acid (TMS)	C12H30O4Si3	322.1452	59559	1001	592	409	96.3325	22.423	19
73	Homovanillic Acid (TMS)	C15H26O4Si2	326.137	58816	1744	875	869	94.3344	21.901	19
74	Fluridone	C19H14F3NO	329.1027	57199	3361	896	2465	91.1605	25.9639	22
75	Fenarimol	C17H12Cl2N2O	330.0327	58670	1890	409	1481	94.6042	18.7164	15
76	Trifluralin	C13H16F3N3O4	335.1093	60005	555	100	455	95.2156	18.6286	16
77	Threonine (TMS)	C13H33NO3Si3	335.1768	59934	626	343	283	93.5062	15.1307	13
78	Cysteine (TMS)	C12H31NO2SSi3	337.1383	60044	516	43	473	95.6321	24.3446	20
79	Ferulic Acid (TMS)	C16H26O4Si2	338.137	58658	1902	833	1069	93.7208	20.5762	18
80	Estrone (TMS)	C21H30O2Si	342.2015	58774	1786	1190	596	95.6687	17.1879	15
81	Trans-4-Hydroxyproline (TMS)	C14H33NO3Si3	347.1768	60138	422	217	205	92.8455	14.7902	13
82	Ornithine (TMS)	C14H36N2O2Si3	348.2085	60235	325	160	165	94.992	16.6606	16

83	Aspartic Acid (TMS)	C13H31NO4Si3	349.1561	60081	479	236	243	95.4653	20.5802	18
84	Glutamine (TMS)	C14H34N2O3Si3	362.1877	60357	203	128	75	95.8571	18.9067	18
85	Glutamic Acid (TMS)	C14H33NO4Si3	363.1717	59782	778	265	513	93.0214	19.4464	17
86	Sinapic Acid (TMS)	C17H28O5Si2	368.1475	57349	3211	516	2695	92.4176	21.7295	19
87	Dopamine (TMS)	C17H35NO2Si3	369.1976	59815	745	325	420	94.1092	13.6762	11
88	Histidine (TMS)	C15H33N3O2Si3	371.1881	60263	297	65	232	96.2284	21.8017	19
89	Orotic Acid (TMS)	C14H28N2O4Si3	372.1357	59701	859	104	755	91.4427	20.3166	17
90	Loratadine	C22H23ClN2O2	382.1448	58320	2240	210	2030	95.5911	23.8813	20
91	Pyroxidine (TMS)	C17H35NO3Si3	385.1925	60013	547	307	240	94.5833	13.25	11
92	Tyrosine (TMS)	C18H35NO3Si3	397.1925	59986	574	280	294	95.3231	13.6224	11
93	Tryptophan (TMS)	C20H36N2O2Si3	420.2085	60117	443	111	332	95.9839	17.6175	14
94	Lysine (TMS)	C18H46N2O2Si4	434.2636	60292	268	37	231	95.9536	19.1255	16
95	Ascorbic Acid (TMS)	C18H40O6Si4	464.1902	60098	462	153	309	94.5365	21.5049	18
96	2'-Deoxyadenosine (TMS)	C19H37N5O3Si3	467.2204	60406	154	20	134	95.1771	21.0448	19
97	Beta-Sitosterol (TMS)	C32H58OSi	486.4257	60362	198	140	58	97.2639	14.0517	13
98	Estriol (TMS)	C27H48O3Si3	504.2911	60141	419	188	231	95.6443	13.4069	12
99	Cystine (TMS)	C18H44N2O4S2Si4	528.182	60182	378	4	374	89.6661	14.7326	12
100	Uridine (TMS)	C21H44N2O6Si4	532.2276	60226	334	20	314	87.1329	7.8822	5
101	Glucose (TMS)	C21H52O6Si5	540.261	59997	563	58	505	89.2621	10.2832	7
102	Inositol (TMS)	C21H52O6Si5	540.261	59946	614	58	556	89.8296	10.4011	7

103	Adenosine (TMS)	C22H45N5O4Si4	555.2549	60394	166	8	158	91.0997	10.1646	7
104	Glucosamine (TMS)	C24H61NO5Si6	611.3165	60276	284	10	274	82.922	4.6934	4
105	Catechin (TMS)	C30H54O6Si5	650.2767	60278	282	10	272	93.6416	8.8272	7
	Average		298.8377	56998.6	3561.35	1946.81	1614.543	93.5741	19.506	16.581

Supplementary Table 3. Shown here are the associated spectral match score, HRF score, and peak count for all extracted spectra in the drug spike-in dataset. All spectra considered contained at least 10 peaks.

Drug Name	***Concentration***	***Spectral Match***	***HRF Score***	***Peak Count***
Nicotine	10 ng	89.82369	99.17881	101
Nicotine	5 ng	89.21242	99.22686	95
Nicotine	2.5 ng	89.2211	99.34258	97
Nicotine	1 ng	89.2658	99.01598	82
Nicotine	625 pg	86.08654	97.86442	68
Nicotine	313 pg	83.82492	99.35862	52
Nicotine	162 pg	85.98935	97.18288	66
Nicotine	80 pg	75.55134	92.77129	34
Cotinine	10 ng	90.87393	99.81463	96
Cotinine	5 ng	91.49133	99.75887	98
Cotinine	2.5 ng	90.26395	99.94532	91
Cotinine	1 ng	85.73789	99.76351	66
Cotinine	625 pg	84.45779	99.91503	57
Cotinine	313 pg	81.61932	100	40
Cotinine	162 pg	78.77733	99.79162	39
Cotinine	80 pg	59.86455	100	23
Amobarbital	10 ng	86.61869	99.69883	85
Amobarbital	5 ng	86.22043	100	70
Amobarbital	2.5 ng	82.61674	99.32243	44
Amobarbital	1 ng	76.55431	99.67943	48
Amobarbital	625 pg	66.17535	99.73096	35
Amobarbital	313 pg	64.85207	100	18
Amobarbital	162 pg	No Spectrum	No Spectrum	No Spectrum
Amobarbital	80 pg	No Spectrum	No Spectrum	No Spectrum
Gluethimide	10 ng	91.73291	100	89
Gluethimide	5 ng	89.60455	99.93778	69
Gluethimide	2.5 ng	84.1814	100	38
Gluethimide	1 ng	88.73444	99.84825	59
Gluethimide	625 pg	78.63416	99.54788	30
Gluethimide	313 pg	77.581	99.3464	31
Gluethimide	162 pg	63.58836	99.43759	17
Gluethimide	80 pg	49.96783	95.58267	12
Methadone	10 ng	66.05668	99.58029	100
Methadone	5 ng	64.20798	99.68237	92
Methadone	2.5 ng	64.03547	99.2299	88
Methadone	1 ng	57.32097	99.69799	63

Methadone	625 pg	59.02508	99.18545	70
Methadone	313 pg	47.20419	98.70877	59
Methadone	162 pg	56.5431	98.75955	54
Methadone	80 pg	41.49079	99.38454	25
Methaqualone	10 ng	84.13078	99.38832	92
Methaqualone	5 ng	87.4992	99.24683	98
Methaqualone	2.5 ng	84.18102	99.64644	89
Methaqualone	1 ng	86.51924	99.51907	89
Methaqualone	625 pg	83.29513	98.77386	82
Methaqualone	313 pg	81.31826	97.85804	66
Methaqualone	162 pg	80.40196	97.09529	84
Methaqualone	80 pg	72.31447	95.20307	41
Scopolamine	10 ng	92.70723	99.82007	87
Scopolamine	5 ng	90.92564	100	79
Scopolamine	2.5 ng	88.18741	100	61
Scopolamine	1 ng	83.65214	99.53964	52
Scopolamine	625 pg	66.42922	100	35
Scopolamine	313 pg	53.5959	97.49234	17
Scopolamine	162 pg	53.45593	98.32571	24
Scopolamine	80 pg	No Spectrum	No Spectrum	No Spectrum
Primidone	10 ng	89.72626	99.78106	66
Primidone	5 ng	88.58776	99.78101	62
Primidone	2.5 ng	84.03984	99.76632	53
Primidone	1 ng	83.67805	99.74081	42
Primidone	625 pg	59.92945	97.64044	24
Primidone	313 pg	52.30685	92.53424	20
Primidone	162 pg	No Spectrum	No Spectrum	No Spectrum
Primidone	80 pg	No Spectrum	No Spectrum	No Spectrum
Loratidine	10 ng	89.57203	99.53398	149
Loratidine	5 ng	92.88445	99.413	151
Loratidine	2.5 ng	87.91399	99.3452	128
Loratidine	1 ng	83.65915	99.45562	86
Loratidine	625 pg	72.5576	99.83844	53
Loratidine	313 pg	59.45031	100	29
Loratidine	162 pg	60.01962	100	34
Loratidine	80 pg	32.68794	100	10