High Peak Density Artifacts in Fourier Transform Mass Spectra and their Effects on Data Analysis

Joshua M. Mitchell[1,2,3], Robert M. Flight[2,3], Qing Jun Wang[2,6], Woo-Young Kang[3], Richard M Higashi[2,3,5], Teresa W-M Fan[2,3,5], Andrew N. Lane[2,3,5], Hunter N.B. Moseley[1,2,3,4,*]

[1]Department of Molecular & Cellular Biochemistry

[2]Markey Cancer Center

[3]Center for Environment and Systems Biochemistry and the Resource Center for Stable Isotope Resolved Metabolomics

[4]Institute for Biomedical Informatics

[5]Department of Toxicology & Cancer Biology

[6]Department of Ophthalmology and Visual Sciences

University of Kentucky, Lexington KY, United States

*Corresponding Author: hunter.moseley@uky.edu

Keywords: Fourier transform, mass spectrometry, artifact, data analysis, metabolomics

Abstract:

Fourier-transform mass spectrometry (FTMS) allows for the high-throughput and high-resolution detection of thousands of metabolites. Observed spectral features (peaks) that are not isotopologues do not directly correspond to known compounds and cannot be placed into existing metabolic networks. Spectral artifacts account for many of these unidentified peaks, and misassignments made to these artifact peaks can create large interpretative errors. Without accurate identification of artifactual features and correct assignment of real features, discerning their roles within living systems is effectively impossible.

We have observed three types of artifacts unique to FTMS that often result in regions of abnormally high peak density (HPD) which we collectively refer to as HPD artifacts: i) fuzzy sites representing small regions of m/z space with a 'fuzzy' appearance due to the extremely high number of peaks present; ii) ringing due to a very intense peak producing side bands of decreasing intensity that are symmetrically distributed around the main peak; and iii) partial ringing where only a subset of the side bands are observed for an intense peak. Fuzzy sites and partial ringing appear to be novel artifacts previously unreported in the literature and we hypothesize that all three artifact types derive from Fourier transformation-based issues. In some spectra, these artifacts account for roughly a third of the peaks present in the given spectrum. We have developed a set of tools to detect these artifacts and approaches to mitigate their effects on

downstream analyses.

Introduction

Fourier Transform Mass Spectrometry (FT-MS) provides high performance in terms of sensitivity, resolution, and mass accuracy simultaneously over many other types of MS commonly used for biomolecular detection and quantification. These combined capabilities provide tangible analytical and interpretative improvements including: i) the ability to resolve isotopologues with identical unit masses but different real masses enabling multi-element isotopic natural abundance correction (H. N. Moseley, 2010) (Carreer, Flight, & Moseley, 2013) in multiple labeling experiments (Yang, Fan, Lane, & Higashi, 2017a); ii) improved assignment accuracy (although often not fully unambiguous assignments) (Kind & Fiehn, 2006); and iii) the detection of low concentration metabolites in the subfemtomolar range (Eyles & Kaltashov, 2004) (Dettmer, Aronov, & Hammock, 2007). In the metabolomics field, these improvements permit more complicated, but more informative experimental designs such as the use of multiple isotope-labeled precursors for use in stable isotope-resolved metabolomics (SIRM) experiments (Yang et al., 2017a). These experiments can provide the information needed to elucidate unknown metabolic pathways (Creek et al., 2012) (Higashi, Fan, Lorkiewicz, Moseley, & Lane, 2014), quantify the relative flux through metabolic pathways (Hiller, Metallo, Kelleher, & Stephanopoulos, 2010), identify multiple metabolite pools (T. W. M. Fan et al., 2012), and identify active metabolic pathways under various cellular conditions

(Sellers et al., 2015a; Sellers et al., 2015b) . In turn, these informational gains enable researchers to build more complete models of cellular metabolism and to better understand life processes, both healthy and pathological, more completely at a mechanistic level, facilitating the identification of potential targets for pharmaceutical intervention (T. W. Fan et al., 2009) and the quantification of differential drug response at a molecular level (Harris et al., 2012).

While the potential benefits of FT-MS are significant, so are the limitations. In addition to the increased cost and complexity of the instruments, the amount of data produced by a single run on an FT-MS instrument can be quite large and when deployed in a high-throughput environment as with all high data capture ‘omics techniques, necessitates the use of automated tools for data reduction, data quality assessment and control, peak (feature) assignment, and other downstream analyses enabling information extraction and interpretation within reasonable timeframes. In particular, the extraction and interpretation of meaningful data from MS spectral often necessitates correct peak assignment. While peak assignment is relatively straight-forward and well-validated in targeted MS experiments, especially when combined with chromatography and/or done in tandem (Ogura, Bamba, & Fukusaki, 2013) (Zhang & Brodbelt, 2004) (Astarita, Ahmed, & Piomelli, 2009), assignment of untargeted MS analyses of non-polymeric biomolecules remains harder to perform in a rigorous, error-limiting manner, even with the advanced capabilities of FT-MS.

In untargeted FT-MS metabolomics approaches, m/z database-based assignment tools such as LipidSearch (Peake, Yokoi, Wang, & Yingying, 2013)

and PREMISE (in-house tool developed within CESB) have been used to assign FT-MS features observed in direct infusion experiments (Lorkiewicz, Higashi, Lane, & Fan, 2012) (Yang et al., 2017a). These database approaches have the advantage of low computational overhead and *a priori* knowledge when the m/z database is tailored to the biological system being studied. However, tailored m/z databases limit discovery – a stated goal of many untargeted analyses (H. N. B. Moseley, 2013), introduce assignment bias and have difficulty disambiguating possible assignments. Also, these assignment approaches are often based on the matching of a single spectral feature to an m/z value in the database, which are statistically error-prone due to a lack of aggregate, cross-validating evidence.

However, the ability to identify true signal is essential for the accurate assignment and interpretation of all MS datasets, including FT-MS datasets, as all analytical techniques have the potential to generate artifactual signals. Artifactual signals are signals not representative of the sample composition but rather arise from instrumental or data processing limitations. Such signals obviously do not represent the underlying biochemistry of a sample, and at best complicate data interpretation and at worst lead to incorrect interpretations. As the scale of experiments increase and untargeted analyses become increasingly prevalent, as is the case with the field of metabolomics (Goodacre, Vaidyanathan, Dunn, Harrigan, & Kell, 2004), the ability to distinguish sample-related signal and artifactual signal becomes increasingly important.

Fundamentally, artifactual MS peaks can be divided into two major types. First are artifactual peaks that result from unexpected ions during spectral

acquisition. For example, the presence of contaminant compounds (Keller, Sui, Young, & Whittal, 2008) (e.g. plasticizers and keratin) and spontaneous chemical reactions during analysis (e.g. molecular rearrangements (McLafferty, 1959)) are just two of many mechanisms by which unexpected ions can be produced and subsequently detected by a spectrometer. In this case, the peak signals represent real analytes; however, these detected analytes are not representative of the sample. The second type of artifactual peaks are those that do not correspond to ions present during spectral acquisition. The cause of these second types of artifacts typically depends upon the mass spectrometry platform on which the spectrum was acquired. Examples for FTMS will be described below. Also, characteristics of real peaks can be artifactual. For example, ion suppression and overloading an ion trap can distort acquired spectral intensities.

The use of the Fourier transformation in the processing of FT-MS raw data introduces new avenues for generating artifactual peaks in resulting FT-MS spectra. This is generally well known, especially for other analytical techniques that utilize the Fourier transform with common FT-specific artifacts like sinc wiggles in NMR (P. J. Hore, 1985) and side lobes in FTIR (Griffiths & Pariente, 1986). However, previous studies regarding FT-MS specific artifacts have focused almost exclusively on “harmonic peaks” (Mathur & O’Connor, 2009) and “sidebands” or “peak ringing” (Miladinović, Kozhinov, Tsybin, & Tsybin, 2012), which do not cover all the artifact types we have observed in our FT-MS spectra. We have observed two additional FTMS artifacts which we have named fuzzy sites and partial peak ringing.

Materials and Methods:

HPD Artifact detection.

Although the three FT-MS artifacts outlined above (ringing, partial ringing and fuzzy sites) have distinct appearances, they all result in regions of high peak density (HPD). These artifacts can be collectively referred to as HPD artifacts due to this shared property and we leverage this property to create an automated tool for their detection. Our HPD-detector is implemented using the Python programming language (Van Rossum & Drake Jr, 1995) version 3.4 and makes extensive use of the Numpy library (Walt, Colbert, & Varoquaux, 2011) to accelerate calculations. Starting with a peaklist in a Javascript Object Notation (JSON) format (Supplemental Figure 2), the detector first parses and sorts the peaks in ascending order of their m/z values, enabling the use of algorithmically efficient methods for searching the peaklist by m/z, namely a binary search. Next, a 1 m/z window is slid across the spectrum in 0.1 m/z increments. At each increment, a pair of binary searches are used to find the peaks within the window that are then counted to give a peak density metric (Step 1 in Figure 1) that is assigned to the mean m/z of the window.

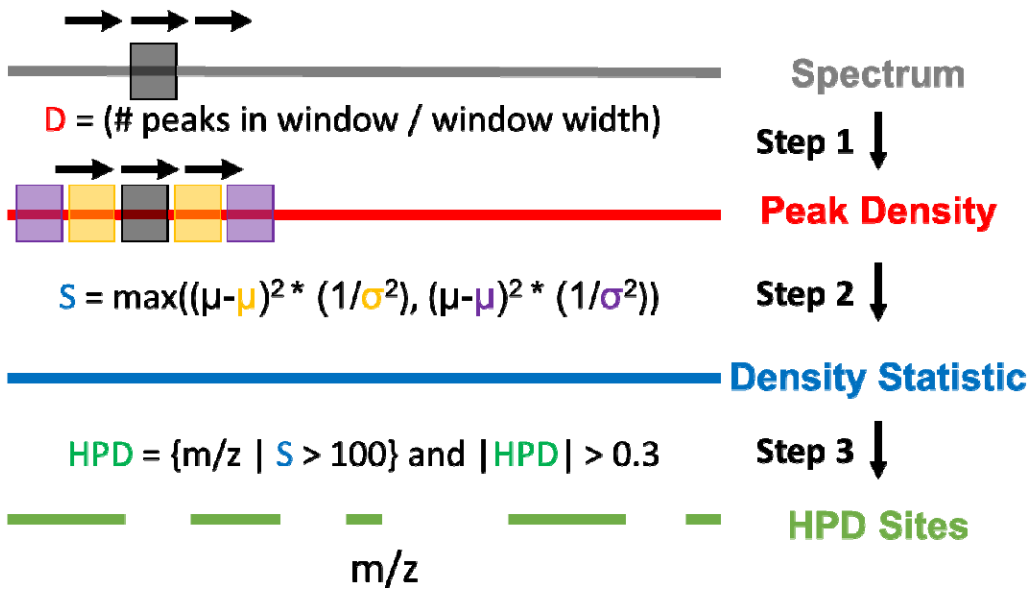


Figure 1: Automated HPD-Site Detection. The HPD-site detection algorithm requires three steps. First, a peak density metric is calculated for the spectrum using a sliding window method. Second, a set of K+1 windows and the peak density metric are used to calculate a density statistic for each portion of the spectrum. This metric flattens out density differences due to signal-to-noise differences or baseline differences and highlights spectra with HPD artifacts (Fig 2E, F, G, H). Filtering this metric reveals the location of the HPD artifacts.

By comparing these peak density statistics to one another, HPD artifacts can be found. This comparison requires applying another window operation to the density statistics calculated previously. In this operation, N pairs of non-overlapping 'reference' windows distributed symmetrically around a single 'test' window are moved across the spectrum in 0.1 m/z increments. The test window is the region of spectrum being tested for HPD phenomena using the reference windows as estimates of "normal" peak densities. At each increment, the mean and standard deviation of each test window and each pair of reference windows are calculated. The mean of the test window is then compared to the mean of each pair of reference windows and normalized by the standard deviation of the corresponding pair of reference windows yielding N chi-squared-inspired

statistics for each test window and the max value is assigned to the mean m/z of the test window (Step 2 in Figure 1). Taking the maximum comparison value across multiple pairs of reference windows increases the sensitivity of the test and insulates the method against the edge case where all reference windows and the test window contain HPD artifacts and thus do not appear statistically different from one another. Although higher values of N are theoretically superior, testing showed no significant improvement for N > 2.

In the final step, the continuous subdomains of m/z space at least 0.3 m/z in width with density statistic values over 100 are reported (Step 3 in Figure 1-). These regions likely contain some form of HPD phenomena as they have significant differences in peak density compared to neighboring regions.

Ringing detection.

The HPD-detector outlined above detects fuzzy sites and some instances of ringing and partial ringing but has lower sensitivity for ringing artifacts. Partial ringing often does not increase peak density sufficiently to be detected by the HPD-detector, while true ringing can often occupy a very wide m/z window, making their detection more difficult. Therefore, we created a dedicated ringing detector in Python 3.4 that identifies sets of peaks that are possibly assignable to the same compound and has matching m/z offsets to other expected compounds.

The ringing detector requires as input a peaklist in JSON a list of expected m/z values in JSON (Supplemental Figure 3) and a description of the adducts to

consider for each ion (Supplemental Figure 4). The provided list of expected m/z values was built using the MaConDa database (Weber, Li, Bruty, He, & Viant, 2012) of common mass spectrometry contaminants and each entry represents the non-adducted monoisotopic form of a contaminant while the adduct and isotopologue file was created manually.

In the first step of the analysis, the ringing detector expands the list of expected m/z values by generating their adducts and a set of expected isotopologues as specified by the adduct and isotopologues file. This expansion ensures that we are searching for the most likely forms of the provided compounds. With the expanded set of expected m/z values calculated, the detector parses the peaklist and sorts the peaks in ascending m/z order to enable binary searching by m/z value. Next, an estimate of the noise intensity level is calculated. From an ascending list of peak intensities, the ringing detector first identifies the first quartile of the intensity values, i.e. the lowest quartile of intensities, and copies them to a noise intensity list. While the standard deviation of the noise intensity list is less than 1.2 times the mean of the noise intensity list, the next 100 intensities are copied to the noise intensity list. Once this process is complete, the median of the noise intensities is calculated. Five times the median noise intensity is considered the intensity cutoff for the spectrum and intensities above the cutoff are very likely “true” signal. A noise sample is then generated by selecting a random set of peaks below the noise cutoff equal in magnitude to the set of peaks above the cutoff.

In this noise sample, the ringing detector searches for the expected m/z values and for every expected m/z in the set of expanded m/z values, a binary search is used to find all peaks within +/-0.05 m/z of the expected m/z value. The m/z difference between each found peak and all other peaks in the noise sample is calculated (peak mass difference vector), as is the m/z difference between the expected m/z and all other expected m/z values (query mass difference vector). The number of matching m/z differences within a match-difference m/z tolerance between the peak mass difference vector and the query mass difference vector are counted. For the ultra-high resolution FTMS spectra analyzed in this study, we used a tight match-difference tolerance of ±0.00005 m/z. The peak with the most matching m/z differences and the number of matches is reported. Since these are noise peaks, these matches are unlikely to be correct and the number of m/z difference matches is due to random chance. Therefore, the ringing detector uses this sample of mostly false matches to estimate a cutoff for accepting true matches. The detector calculates this cutoff as the mean of noise matching count plus three times the standard deviation of the distribution of noise matching counts.

Finally, the ringing detector repeats the search procedure using the entire peaklist. During this search, only peaks with many matching m/z differences greater than the match cutoff calculated from the noise peaks are reported as assignments. When ringing or partial ringing occurs, multiple assignments occur for an expected m/z value.

FT-MS instruments.

To determine the instrument dependence of various artifacts described here, spectra from several different FT-MS instruments were used. The first set of instruments were three Thermo Tribrid Fusion instruments. Fusion 1 and Fusion 2 are maintained by the Center for Environmental and Systems Biochemistry (CESB) at the University of Kentucky, while Fusion 3 (Serial # FSN 10144) is maintained by National Resource for the Mass Spectrometric Analysis of Biological Macromolecule at the Rockefeller University. Fusion 1 (Serial # FSN10115) was delivered to CESB in October 2013. In March of 2016, Fusion 1 had its firmware upgraded to the most recent version at that time. Pre-firmware upgrade Fusion 1 (Fusion 1 – Before) has different HPD artifact patterns than post-firmware Fusion 1 (Fusion 1 – After). Fusion 2 (Serial # FSN10352) was delivered to CESB in May 2015 with the upgraded firmware. Spectra from a Thermo Scientific Orbitrap Fusion Lumos Tribrid mass spectrometer (Serial # FSN20208, delivery data July 2016), which is maintained by the Proteomics Resource Center at the New York University Langone Medical Center, were also examined. The Lumos represents an improvement upon the original version of the Tribrid Fusion. In addition to the Fusion instruments, we examined spectra from a Thermo Q-Exactive+ instrument, which is another orbitrap instrument maintained by High Resolution Metabolomics Laboratory (HRML) at the Institute of Biological, Environmental and Rural Sciences. Also, spectra from a Bruker Solarix instrument (Serial # 150506 A), an ICR-type FT-MS delivered to CESB in April 2014, were examined.

Samples analyzed by FT-MS.

Sample A: Solvent Blanks with and without Avanti Lipid Standards

Solvent blank was composed of Isopropanol:MeOH:Chloroform 800µl:344µl:200µl. Solvent blank was mixed with 28µl 1 M ammonium formate (final ~20 mM; Aldrich #516961), and without or with 70µl Avanti SPLASH™ Lipidomix® Mass Spec Standard (cat# 330707) in MeOH. Solvent blank without or with lipid standards was loaded onto a 96-well polypropylene PCR plate (USA Scientific cat# 1402-9800) and 15 µl was injected into the Fusion 1 by direct infusion through an Advion nanomate. Various resolution and microscan settings were tested for the orbitrap mass analyzer in positive mode with 7 min acquisition, normal mass range between 150-1600 m/z, S-lens RF level 60%, AGC target 1e5, maximum injection time 100 ms, and Easy-IC on.

Sample B: Mouse Liver ICMS Standard

Mouse livers were excised from NSG mice within 5 minutes for euthanizing, and flash frozen in liquid nitrogen and ground into powder under liq. N_2 to <10 um particles. Approximately 0.5 g of the powder was extracted with 50 mL Acetonitrile: water (6:4, v/v). After centrifugation at 22 k rpm and 4 ˚C for 20 min, the supernatant containing polar extracts was distributed into aliquots and lyophilized for long-term storage at -80 ˚C. Immediately before injection, the lyophilized powder was reconstituted with water and 10 µL was injected onto an

ICS5000+ system (Dionex) interfaced to the FT-MS (Fusion 2). Data were acquired in negative mode at a resolving power of 500,000 (at m/z=200) over 52 min of chromatography. The mass range was set between m/z 80-700, maximum injection time was 100 ms with 1 microscan, AGC target was 2e5, S-lens RF level was 60%, and Easy-IC was turned on for internal mass calibration.

Sample C: ECF Solvent Standard

The ECF solvent blank was composed of acetonitrile: water 9:1 (v/v) with a concentration of 20 μM NaCl to convert positively charged ions into sodium adducts (Yang et al., 2017a; Yang, Fan, Lane, & Higashi, 2017b).

Sample D: Paired Human NSCLC Cancer and Non-Cancer Tissue Samples

The samples analyzed are the lipids extracted from flash frozen resected lung tissues from human subjects with resectable stage I or IIa NSCLC (non-small cell lung cancer) collected under a University of Louisville approved Internal Review Board (IRB) protocol. Paired cancerous and non-cancerous tissue 5 cm form the tumor margins in the same lobe of the lung were cut by the surgeon and immediately flash frozen in liq. N_2, and stored at <-80°C prior to metabolite extraction as previously described (Sellers et al., 2015a). Samples were pulverized under liq. N_2 to <10 um, and extracted using a modified Folch method as previously described (Ren et al., 2014) 1 mM butylated hydroxytolune was added to the lipid layer and then dried by vacuum cetrifugatin at room temerpaure. Samples for FT-MS analysis were redissolved in

isopropanol/methanol/chloroform 4/2/1 with 20 mM ammonium formate (95 µl of solvent, 5 µl of sample).

Mass Spectrometry Analysis of Samples A through D

Ultrahigh resolution (UHR) mass spectrometry was carried out on a Thermo Orbitrap Fusion interfaced to an Advion Nanomate nanoelectrospray source using the Advion "type A" chip, also from Advion, inc. (chip p/n HD_A_384). The nanospray conditions on the Advion Nanomate were as follows: sample volume in wells in 96 well plate – 50 µl, sample volume taken up by tip for analysis – 15 µl, delivery time – 16 minutes, gas pressure – 0.4 psi, voltage applied – 1.5 kV, polarity – positive, pre-piercing depth – 10 mm. The Orbitrap Fusion Mass Spectrometer method duration was 15 minutes, and the conditions during the first 7 minutes were as follows: scan type – MS, detector type – Orbitrap, resolution – 450,000, lock mass with internal calibrant turned on, scan range (m/z) – 150 – 1600, S-Lens RF Level (%) – 60, AGC Target – $1.0e^5$, maximum injection time (ms) – 100, microscans – 10, data type – profile, polarity - positive. For the next 8 minutes, the conditions were as follows for the MS/MS analysis: MS properties: detector type – Orbitrap, resolution – 120,000, scan range (m/z) – 150 – 1600, AGC Target – $2.0e^5$, maximum injection time (ms) – 100, microscans – 2, data type – profile, polarity – negative; monoisotopic precursor selection – applied, top 500 most intense peaks evaluated with minimum intensity of $5e^3$ counts; data dependent MS^n scan properties: MS^n level – 2, isolation mode – quadrupole, isolation window (m/z) – 1, activation type – HCD, HCD collision energy (%) – 25, collision gas – Nitrogen, detector –

Orbitrap, scan range mode – auto m/z normal, Orbitrap resolution – 120,000, first mass (m/z) – 120, maximum injection time (ms)-500, AGC target – $5e^4$, data type – profile, polarity – positive. The ion transfer tube temperature was 275 ^oC. (Yang et al., 2017a, 2017b).

Sample E: Human Plasma

Pooled lithium heparin treated plasma (Seralab) was extracted using the methods described by Koulman et al, 2017 (Acharjee et al., 2017). Briefly, 15 uL of plasma was extracted with 100 uL of ultra-pure H2O in a glass vial (2 mL). 250 uL of MeOH was added, and lipids were partitioned into 500 uL of Methyl-tertiary-butyl ether. Following centrifugation (13,000 rpm, 4 degC, 4mins), a 20 uL aliquot of the organic layer was then transferred to a 96-well glass coated plate (ThermoFisher). 95 uL of a solution containing 7.5mM NH4Ac in IPA:MeoH [2:1] was also added to the well. Direct infusion high-resolution mass spectrometry was performed using on a Q-Exactive+ Orbitrap (Thermo), equipped with a Triversa Nanomate (Advion). The Nanomate infusion mandrel was used to pierce the seal of each well before analysis, after which, with a fresh tip, 5 µL of sample was aspirated, followed by an air gap (1.5 µL).

Results

Manual Investigation of Artifacts

Before developing automated tools for their detection, many spectra were manually inspected and the observed artifacts were characterized. Three major artifact types were observed with the HPD property (Figure 2): fuzzy sites, peak ringing and peak partial ringing (Figure 3) and all three appear distinct at the aggregate and scan level. All three artifacts are obvious upon manual inspection, but the growing popularity of FT-MS-based experiments necessitates automated methods for their efficient identification.

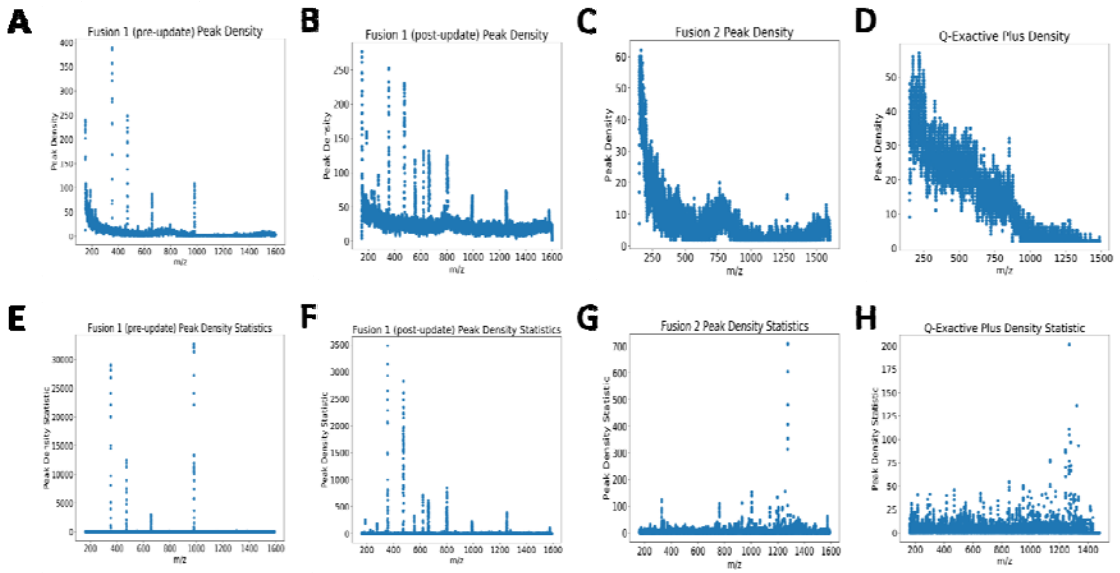


Figure 2: **Peak Density and Peak Density Statistics.** Peak density metric plots produced by our HPD-detector tool highlight the impact of the instrument on peak density and HPD artifact location. All instruments have higher peak densities at lower m/z representing trends in signal-to-noise with respect to m/z in FT-MS. The sharp spikes in peak density correspond to HPD artifacts. The location of these spikes on Fusion 1 is different before and after the firmware update (A, B), suggesting instrument-level processing of the data is related to HPD generation. E-H show the effectiveness of our peak density statistic metric for flattening the non-constant baseline observed in plots of the raw peak density. Without this correction, identifying HPD regions reliably is difficult. The presence of HPD regions in effectively all of our tested Fusion spectra contrasts with the lack of any severe HPD results in Q Exactive spectra supports an instrument level explanation for HPD phenomena. A, B, C, E, F, G were generated from spectra acquired using sample C. D and H were acquired using sample E.

General HPD detection across FT-MS instruments.

Using the HPD artifact detector, we generated plots of peak density for a variety of example spectra across various FT-MS instruments (Figure 2). For all Orbitrap instruments, the peak density decreases monotonically with respect to m/z. While the same general decreasing trend is observed in our ICR instrument, and a Lumos instrument the overall curve is not monotonic (Supplemental Figure 1). These trends are partially explained by differences in signal-to-noise ratio at different m/z values and between different instruments. The peak density properties demonstrate that a statistical approach is necessary to identify HPD as there exists no sufficiently sensitive and selective cutoff on raw peak density for all m/z or all instruments. However, our statistical approach compensates for changes in baseline peak densities revealing regions of significantly higher peak densities (Figure 2D-H).

Although fluctuations in peak densities are expected due to differences in the distribution of compounds in m/z space, this fails to explain the massive difference in peak density statistics present in the spectra and similar patterns are observed in blanks and in scans from failed injections, which resulted in little analyte or solvent. Furthermore, analytical samples derived from the same biological sample have HPD regions at different locations with different instruments and the location of these artifacts differs before and after a firmware update on the Fusion 1 machine (Figures 2A, 2B, 2E, 2F). Manual inspection of a subset of the detected HPD regions consistently failed to identify patterns between peaks that are explainable by chemical phenomena (*e.g.* isotopologues,

different charges, etc.). Together, these findings support an artifactual basis for these regions of spectra and suggest an instrument-level effect leading to their production.

Detection and characterization of fuzzy sites.

The fuzzy site artifacts were first observed and described as “ugly sites”. At the aggregate spectrum level (Figure 3A, 4A,C), fuzzy sites have HPD and a pseudo-Gaussian distribution of peak intensities between the noise baseline and presumed signal peaks. The intermediate intensities of these peaks make identifying and filtering these regions by intensity alone difficult. Fuzzy regions, like other HPD artifacts, have peak m/z differences that are not explainable by isotopologue, charge, or harmonic patterns. A typical fuzzy site occupies a small m/z window, with larger ranges typically occurring at higher m/z values. Fuzzy sites rarely occur singly: many fuzzy sites are typically observed in a single spectrum. Collectively, these sites can represent a significant portion of the peaks but a smaller amount of the total m/z range. Fuzzy site location varies between analytical replicates on the same instrument and with sample composition (Supplemental Figure 6). Fuzzy sites have been observed in samples with failing or no injection as well.

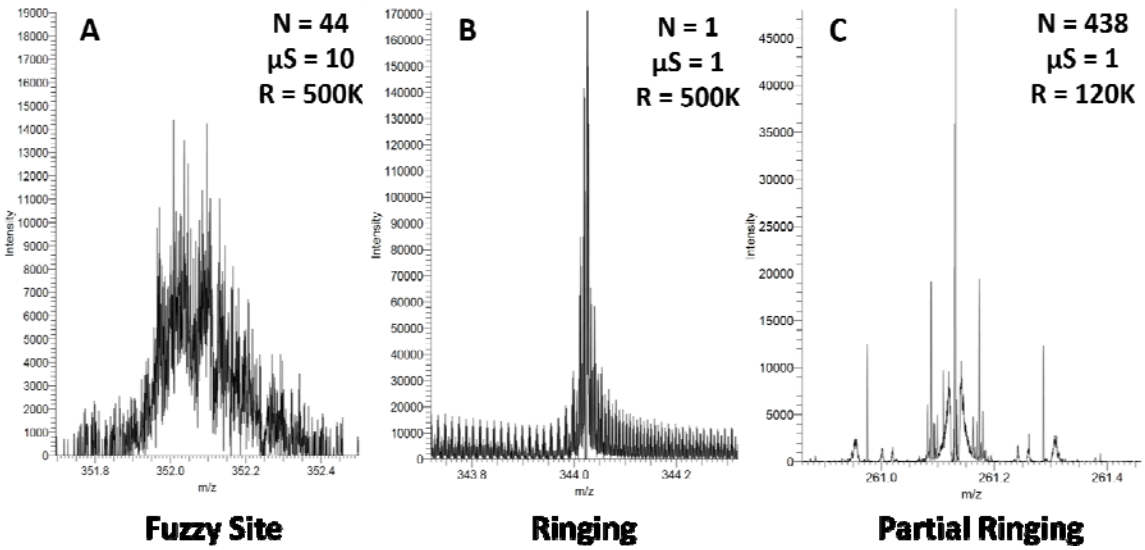


Figure 3: Three Types of HPD Artifacts. There are three subclasses of HPD artifacts we have investigated. The first is the fuzzy site which we believe is a novel artifact type (A, Sample A). A fuzzy site appears pseudo-Gaussian at the aggregate level. Second is ringing, a well-known FTMS artifact where a single intense peak has many side peaks (B, Sample B). We only observed ringing at the scan level. The third artifact is partial ringing which is a ringing-like artifact at the aggregate level (C, Sample C). Partial ringing appears like ringing but has not been discussed in the literature.

Fuzzy sites also have interesting properties at the scan-level as well. While the timing between scans and injection as well as inconsistencies between injections can result in non-perfect scan-to-scan correspondence between peaks (e.g. a peak is present in scan X, but not in scan Y), peaks of the same chemical origin should appear consistently between scans near their true m/z roughly within the resolution of the instrument. At the scan level, the peaks in a mass range identified as a fuzzy site at the aggregate level have very low peak correspondence (Figure 4B). In any given scan, only sections of the fuzzy site region will have peaks and those sections that are populated with peaks change from scan to scan. However, as increasingly more scans are averaged together,

the Gaussian-like distribution of a fuzzy site at the aggregate level becomes clearer (Figure 4A,C). Fuzzy sites appear distinct from either peak ringing or partial peak ringing and represent a novel class of artifact not previously described in the FT-MS literature.

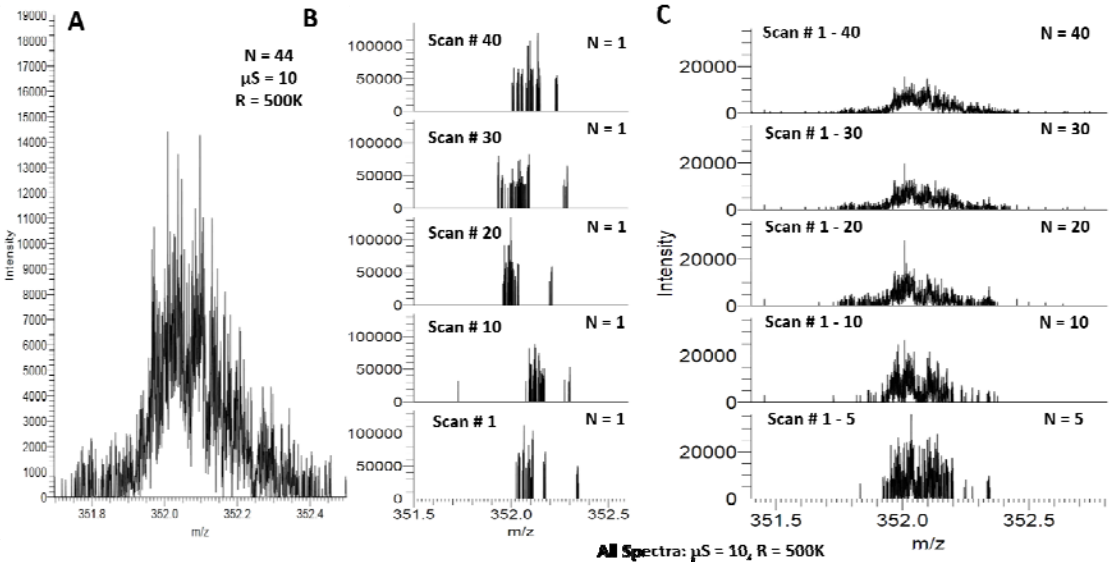


Figure 4: Fuzzy sites at the Aggregate and Scan Level. A typical fuzzy site (A) occupies .5 to 2 m/z at the aggregate level and has a distinct 'fuzzy' appearance due to very high peak density (this image is identical to 3A). At the scan level, only a subdomain of the m/z occupied by the fuzzy site contains peaks; the subdomain that is occupied varies from scan-to-scan (B). As increasingly more scans are aggregated together, the peak distribution converges to the pattern observed at the aggregate level (C). All panels were generated using Sample A.

The initial studies into fuzzy site HPD artifacts were performed on our Fusion 1 instrument where they were first observed. After developing our tools on this data, spectra from other instruments were examined for fuzzy sites to determine if these artifacts were limited to only our instruments. Spectra from the Fusion 2 instrument also contained fuzzy sites as did spectra from Fusion 3. To date, we have observed fuzzy sites in spectra from every non-Lumos Tribrid

Fusion instrument examined. However, we did not find fuzzy sites in spectra examined from any other type of FTMS instrument.

Fuzzy site characteristics varies with resolution and microscan settings.

With fuzzy sites found in all tested non-Lumos Tribrid Fusion spectra but not spectra from other instruments used in our study, we began investigating the effect of two instrument parameters on the appearance of these artifacts. The first parameter examined was resolution, which was a parameter that was improved by the firmware update that changed the HPD properties of Fusion 1. The most up-to-date Fusion has a maximum resolution of 500K at 200 m/z (450K before update) and the second parameter was the number of microscans per scan. The number of microscans is the number of FIDs acquired and summed to create the FID that is transformed to produce the scan-level spectrum. Due to the high scan-level variability of fuzzy sites, these settings were of interest as it directly impacts how scan-level FIDs are acquired and processed.

Using a series of analytical replicates created using the same sample of solvent blank with lipid standards, spectra were acquired on the Fusion 1 instrument at 3 resolutions (120 k, 240 k and 500 k at m/z=200) and 4 microscan settings (1,2,5, and 10 microscans, but only 1,5, and 10 microscans are shown in Figure 5). No combination of settings eliminated the fuzzy sites, but they do change the general appearance of the fuzzy sites. Higher resolution results increased the peak density of fuzzy sites, indicating that the peaks within these regions may be sharper (smaller peak widths) than what is indicated at the

highest resolution. Higher microscan settings increased the variability in peak intensities. Also, extremely high microscan settings (e.g. µS = 350) resulted in broad uniform regions for these fuzzy sites (Supplemental Figure 5).

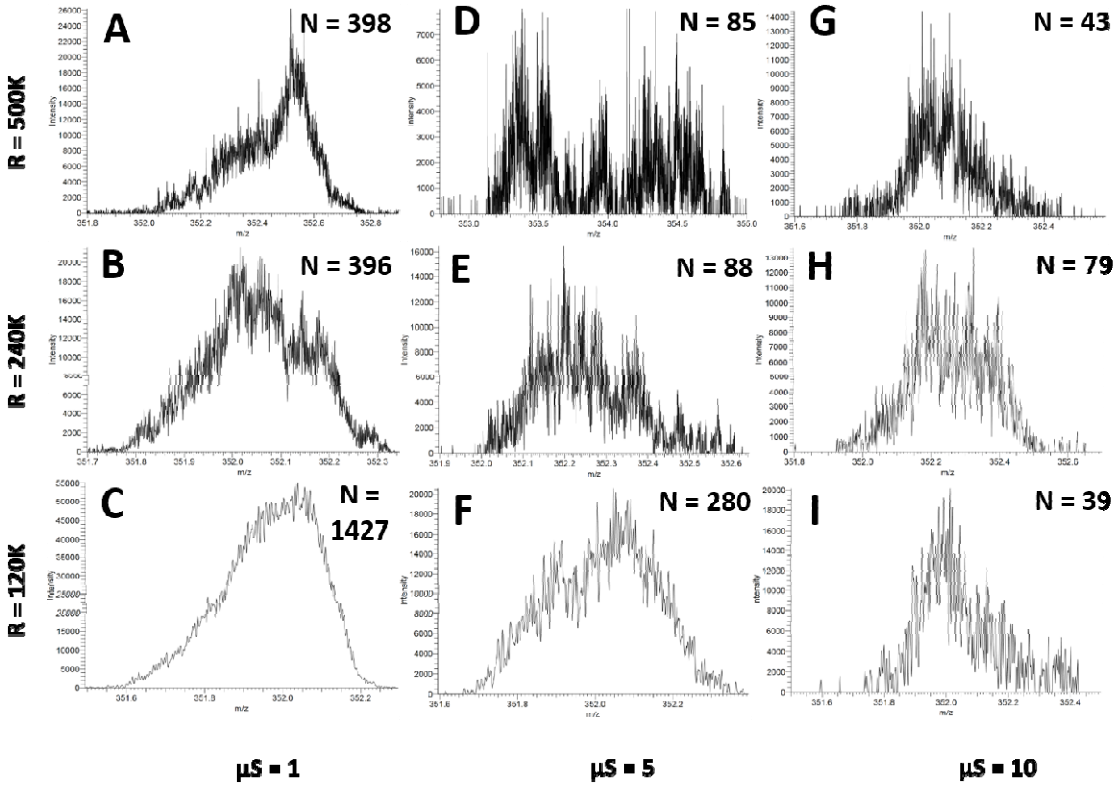


Figure 5: Effect of Resolution and Microscan (µS) on Fuzzy Sites. Permuting over multiple resolution and µS settings shows that no combination of tested settings eliminated fuzzy sites but these settings do change their appearance (A-I). Number of scans collected were set so that total acquisition time was constant. Higher µS increases intensity variance with minimal impact on peak density. Increasing resolution increases peak density but has a lesser impact on peak intensity variance. All panels were generated using Sample A.

Fuzzy site locations are sample specific, confounding these artifacts with sample class

As shown in Supplemental Figure 6, fuzzy site location varies between spectra and appears to shift significantly (a shift greater than the resolution of the

instrument) with changes in sample composition. A potential hypothesis explaining this observation is that sample composition is related to fuzzy site location, possibly due to automatic optimization of instrument parameters.

If this hypothesis is correct, fuzzy site artifacts are a potential problem for real biological applications of FT-MS. To illustrate this potential problem, consider sample D, the paired cancer and non-cancer lung tissue slices. Due to the differences in the concentrations of various metabolites between cancer and non-cancer, we would anticipate that features assigned to spectra of different sample classes could be used to distinguish sample classes. However, if fuzzy sites vary with sample class as well, artefactual features will also distinguish sample class without reflecting the underlying biochemical differences between the classes directly.

This creates a potential confounding factor in all downstream statistical analyses and reduces robustness of acquired spectra if spurious changes in sample conditions introduce new sample-specific artifacts. For example, machine learning methods such as random forest (Breiman, 2001) are trained with known classes of samples in order to classify unknown samples into one or more known classes based on spectral features identified in the training set of samples. However, these techniques rely upon an important assumption that detectable artifacts are not confounded with sample class, which may not be true. The large number of sample-specific artifactual peaks produced by these fuzzy sites can hijack the classifier training, anchoring the classification to artifacts that may change due to unforeseen sample conditions in unknown

samples or changes in the analytical instrument. This effect is demonstrated in Table 1 and can greatly reduce the robustness of the classification of other samples analyzed (Table 1A). In addition, it is non-trivial to remove HPD features correctly. If a feature within an HPD region in any spectrum is just removed, this actually encodes the HPD region due to the absence of a peak (Table 1B). Therefore, the proper action is to remove the HPD-tainted feature from all spectra (Table 1C).

Table 1: Fuzzy Sites are Class Specific. 20 random forests were trained using lipid search assignments from sample D and the mean importance of each feature was reported. HPD features in red were within an HPD region in at least one sample. (A) Without HPD feature removal, on average 6 out of the top 30 features are HPD features . (B) Simple removal of HPD features only within the specific HPD region encodes the HPD region into the features, which falsely boosts its importance. (C) The proper action is to remove that feature occurring within HPD regions across any spectra from all spectra. All three classifiers could disambiguate cancer and non-cancer with 100% accuracy.

A

No HPD Removal		
Feature Name	Mean Importance	HPD?
X350	0.025284079	FALSE
X1009	0.017746561	FALSE
X349	0.016622113	FALSE
X402	0.01461637	FALSE
X431	0.014193132	FALSE
X723	0.013454143	FALSE
X347	0.013248217	FALSE
X401	0.012996595	FALSE
X736	0.01271066	TRUE
X742	0.012631245	TRUE
X734	0.012398963	FALSE
X729	0.012233557	FALSE
X2082	0.012101343	TRUE
X726	0.011743838	FALSE
X2054	0.01170841	FALSE
X404	0.011433176	FALSE
X2102	0.011063981	FALSE
X701	0.01049329	FALSE
X743	0.009800354	TRUE
X732	0.009631989	FALSE
X11644	0.008841659	TRUE
X700	0.008828759	FALSE
X162	0.008565117	FALSE
X9061	0.008523628	FALSE
X2056	0.008452536	FALSE
X2085	0.008405387	FALSE
X2055	0.008376826	FALSE
X373	0.008356794	FALSE
X1506	0.008204027	TRUE
X318	0.008176458	FALSE

B

Encoding Removal		
Feature Name	Mean Importance	HPD?
X350	0.027541747	FALSE
X1009	0.017562844	FALSE
X349	0.017010815	FALSE
X402	0.016891887	FALSE
X431	0.015439427	FALSE
X347	0.014695127	FALSE
X723	0.014389477	FALSE
X726	0.01338201	FALSE
X734	0.013149325	FALSE
X2054	0.013108529	FALSE
X736	0.012994823	TRUE
X2102	0.012842074	FALSE
X404	0.012733523	FALSE
X701	0.011989011	FALSE
X729	0.011986801	FALSE
X732	0.010204682	FALSE
X162	0.00919584	FALSE
X715	0.009022132	FALSE
X700	0.008776501	FALSE
X1348	0.008602804	FALSE
X373	0.008574812	FALSE
X2056	0.008477637	FALSE
X2055	0.008347384	FALSE
X365	0.008166483	FALSE
X318	0.00811912	FALSE
X2084	0.007900575	FALSE
X706	0.007858253	TRUE
X704	0.007857503	FALSE
X375	0.007825088	FALSE
X405	0.007816595	FALSE

C

Non-Encoding Removal		
Feature Name	Mean Importance	HPD?
X350	0.029486	FALSE
X1009	0.020354687	FALSE
X349	0.019596533	FALSE
X402	0.017925542	FALSE
X431	0.017737243	FALSE
X723	0.016257434	FALSE
X347	0.015685867	FALSE
X734	0.014918165	FALSE
X726	0.014901981	FALSE
X404	0.014534893	FALSE
X729	0.014501878	FALSE
X2054	0.013354495	FALSE
X2102	0.013171222	FALSE
X701	0.012675722	FALSE
X732	0.01151158	FALSE
X162	0.010887918	FALSE
X715	0.010324749	FALSE
X700	0.009990557	FALSE
X2055	0.009950157	FALSE
X318	0.009572078	FALSE
X2056	0.009484001	FALSE
X2085	0.009065236	FALSE
X1348	0.008925486	FALSE
X275	0.00883647	FALSE
X373	0.008730355	FALSE
X954	0.008564601	FALSE
X1007	0.008476491	FALSE
X405	0.008432963	FALSE
X430	0.008412441	FALSE
X704	0.008375828	FALSE

Peak ringing.

Peak ringing is an HPD artifact characterized by the presence of many peaks symmetrically centered around a very intense primary peak. The intensity of these side peaks decreases with increasing distance from the primary peak and the m/z interval between each side peak is consistent across the entire artifact. This pattern is clear in Figures 3B and 6A. Unlike fuzzy sites, where artifactual peaks appear and disappear from scan-to-scan, peak ringing is an all-or-none phenomenon at the scan level (Figure 6A); however, ringing peaks may not ring in all scans (Figure 6B) and may exhibit other artifact types (Figure 6C, D). In general, only peaks with high relative intensity in a scan will exhibit ringing behavior. Peak ringing is a well-known artifact type in FT-based instruments and can be caused by Fourier transformation of a truncated FID (Wood & Mark Henkelman, 1985). We hypothesize that this is the cause in our examples as well. There are well-known solutions to suppressing these artefacts using processing in the time domain (Guan & Marshall, 1997).

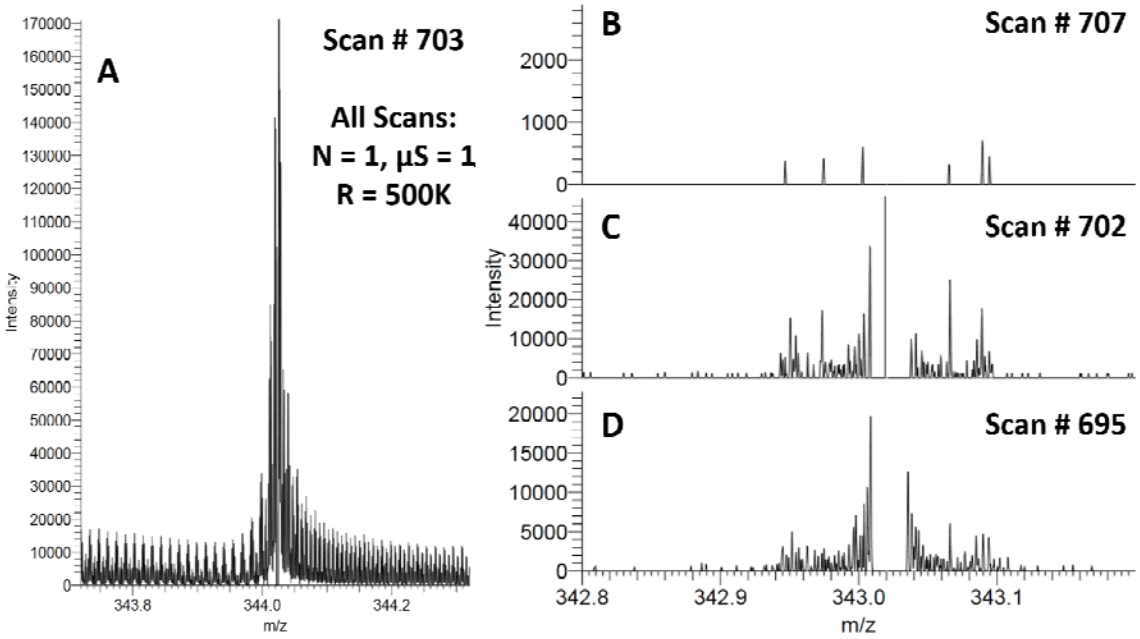


Figure 6: Ringing Artifacts. Ringing results in many artifactual peaks for an intense peak at the scan-level with a distinctive pattern of decaying intensity with increasing distance from the central peak (A). A complete complement of ringing peaks is present for a primary peak in a scan with true ringing. A peak that rings in one scan, does not necessarily ring in other scans; some scans demonstrate partial ringing (C, D) while other scans show no ringing (B). All panels were generated using Sample B.

Peak partial ringing.

Like ringing, partial ringing occurs around intense peaks (*i.e.*, primary peaks) (Figure 7A) but much like fuzzy sites, the artifactual peaks vary significantly at the scan-level. The artifactual peaks appear symmetrically centered around the primary peak (Figure 7A), occupying several tenths of an m/z, but they rarely occur in the immediate vicinity of the primary peak (Figure 7B, C). Side peaks from partial ringing are of lower intensity than the primary peak, but their intensities do not necessarily decrease with increasing distance from the primary peak. Although the symmetry of the peaks and the intensity pattern is less apparent in some scans (Figure 7D), when sufficient scans are aggregated,

partial ringing appears smoother and more akin to true ringing except near the main peak (Figure 7A). We speculate that partial ringing is the result of incomplete ringing suppression at the scan level.

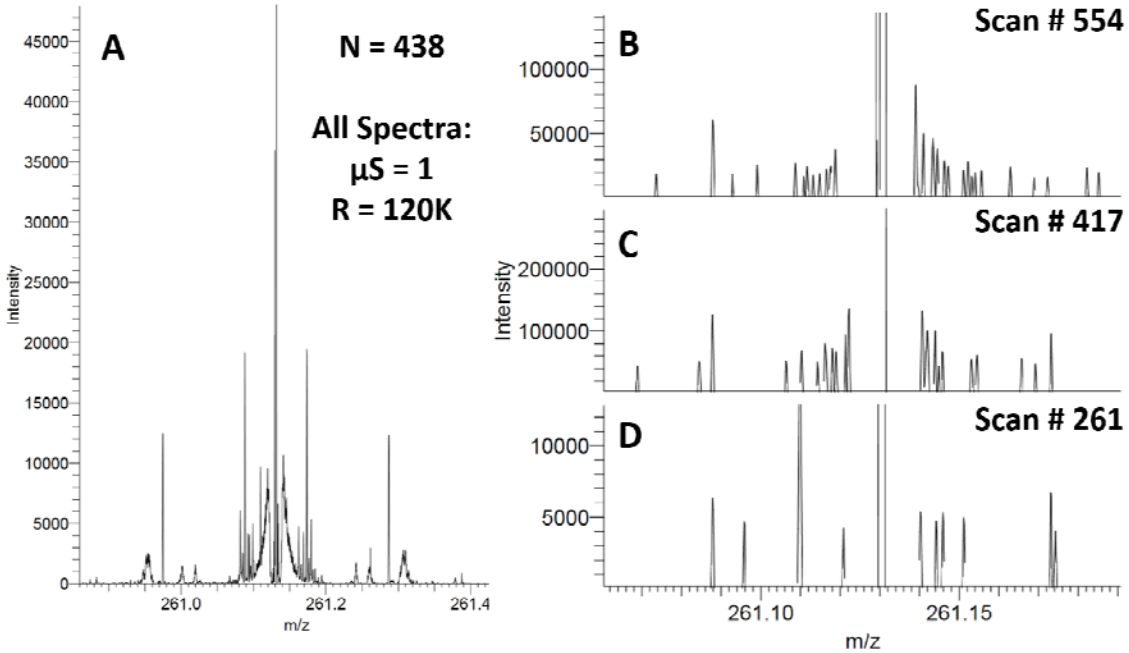


Figure 7: Partial Ringing. (A) Partial ringing produces peak patterns at the aggregate-level that are similar to ringing at the scan-level. Unlike ringing, partial ringing does not strictly decrease with increasing distance from the main peak and partial ringing is often absent (or greatly diminished) near the main peak . (B,C, D) At the scan level, the location of the artifactual peaks is highly variable . All panels were generated using Sample C.

Discussion:

Origin of FT-MS artifacts

FT-related artifacts are ubiquitous in essentially all analytical techniques that leverage FT to process data. A common cause of ringing artifacts (Fig. 3B & 6) is the transformation of a truncated FID and/or improper apodization and these artifacts have analogs in both FT-NMR (*i.e.* sinc-wiggles (P. Hore, 1985; P. J. Hore, 1985)) and FT-IR (*i.e.* sidelobes (Philip B Tooke, 1989) (Herres &

Gronholz, 1984)) . The classic experimental reason is due to insufficient acquisition time combined with inadequate data processing, which results in the Fourier transformation of a truncated FID. In our analysis, ringing was the least commonly observed artifact and was only observed in IC-MS spectra. Under these conditions, there are limitations on FID acquisition time and the separation of compounds by IC results in some scans observing only a small number of compounds with high relative concentrations. In discussions with the instrument vendor (ThermoFisher Scientific), ringing artifacts in our Orbitrap instruments are due to a rare failure in processing software to handle an edge case when the minimum acquisition frequency approaches the Nyquist frequency under a specific range of injection times. Thus, the edge case condition results in a software failure that mimics a classic Fourier transform truncation artifact.

Partial ringing is more difficult to explain; however, the observation of peaks that exhibit ringing in one scan and partial ringing in other scans indicate that the two phenomena are most likely related. Also, partial ringing itself rarely occurs next to an intense peak, but rather in relative vicinity of an intense peak. Both observations suggest that partial ringing may result from ringing suppression on the instrument. The exact mechanism by which ringing is suppressed remains unclear without access to the raw time domain data acquired by the instrument and more detail on the algorithms used to process and transform this data. From discussions with ThermoFisher Scientific, these artifacts may be experimentally due to amplitude modulation of the FID that could have an electronic or vibrational origin.

Scan-level inconsistency of fuzzy sites and the absence of chemical phenomena with explanatory m/z patterns strongly suggests an artifactual origin for fuzzy sites. We hypothesize that fuzzy sites are another form of Fourier transformation artifact; specifically, they are the mass spectrometry equivalent to Gibbs phenomena, where transformation of an FID containing a discontinuity results in the production of many oscillations in a small subdomain of the transformed space. These oscillations would result in many peaks in a small m/z window, which is consistent with the scan-level appearance of fuzzy sites. At first glance, this scan-level manifestation appears inconsistent with the pseudo-Gaussian appearance of fuzzy sites at the aggregate level. This discrepancy can be explained by the central limit theorem. As increasingly more scan-level fuzzy sites are aggregated together around an expected point or region, which is essentially equivalent to averaging multiple small distributions together, the aggregate will converge to a larger Gaussian-like distribution.

The scan-level variability in fuzzy site location may result from changes in the population of ions present in the instrument when a scan is acquired which in turn could change where such discontinuities occur. Discerning the exact cause of these discontinuities is difficult to do as the set of FIDs on the Fusion instruments are not available to the end user. We speculate that discontinuities could be introduced by, but not limited to: improper apodization (or lack thereof) and other data transformations, insufficient acquisition time, hardware-specific problems, or a combination. In discussions with ThermoFisher Scientific, background electronic signals from accidentally oscillating electronic components

either inside the instrument or from outside sources could give rise to phenomena that we detect as fuzzy sites in our non-Lumos Fusion spectra. These signals would appear to “hop” around from scan to scan since they are fundamentally unstable with respect to frequency and amplitude. Furthermore, a range of experimental conditions could influence this hopping, including proximity, temperature and other atmospheric conditions, and sample ion complexity.

Fuzzy sites reduce robustness of downstream data analyses

This hypothetical possibility of impacting data analysis is demonstrated in the following classification results. In this example, spectra ranging from 150 to 1600 m/z were acquired from paired cancer and non-cancer lung tissue sample non-polar extracts in an approximately random order with respect to sample class (see Supplemental Figure 7). Random forest was then used to train a classifier that would classify the samples based on LipidSearch peak assignments to expected lipids. The resulting classifier effectively classifies the samples into the appropriate class; however, these classifiers make extensive use of peaks that are present within fuzzy sites. Although these classifications are accurate, the classification is heavily based on artifactual features without direct molecular interpretation. Also, these same issues would have significant impact on differential abundance analyses as well.

Mitigating the effects of fuzzy sites on downstream data analyses

The fuzzy-site/sample-class confound can be mitigated by simply removing fuzzy site assignments from the assignment lists prior to training the classifiers. However, care must be taken when performing this operation as a sample-by-sample removal of these sample class specific features can encode artifactual class information into the peaklist. For example, if the cancer class of spectra has a fuzzy site from 500 to 502 m/z but not the non-cancer, removal of and therefore absence of spectra from 500 to 502 m/z is also cancer-specific and therefore artifactually usable by machine learning methods. One solution is to remove the union of the m/z regions of all fuzzy sites from all spectra in an experiment. This prevents the accidental absence encoding of class information into the data, but has the disadvantage of removing larger regions of spectrum as the number of samples and classes increases. Table 1 shows the effect of fuzzy site removal on feature importance in classification. Twenty random forests were trained on LipidSearch assignment features derived from Thermo Tribrid Fusion FTMS spectra collected on cancer and non-cancer lung tissue samples (Sample C). The top features based on mean importance of each feature across the 20 random forests are reported. Features in red came from the HPD region of a fuzzy site in at least one sample (HPD features). Without HPD feature removal, on average 6 out of the top 30 features are HPD features. But as mentioned earlier, it is non-trivial to remove HPD features correctly. If a feature is within an HPD region in any spectrum, the proper action is to remove that feature from all spectra, else these features may still contribute to classification. In the latter case, the artificial absence of that feature in some spectra can falsely boost its

importance in the spectra that still contain that feature ad indicated by the middle encoding removal results in Table 1. But when HPD features are removed in a non-encoding manner, no HPD feature makes it into the top 30 mean importance list (as expected). In this example, perfect disambiguation of cancer and non-cancer was possible with and without proper HPD removal; however, proper HPD feature removal is crucial to ensuring that important features represent true biological variance between the sample classes. Additionally, this demonstrates that classification accuracy is not necessarily an indication of classifier quality or robustness, especially when artifacts are present.

Conclusions:

With our HPD detector and ringing detector tools, we have identified and characterized three distinct types of artifacts that produce large numbers of peaks: fuzzy sites, ringing and partial ringing. In some cases, these artifacts can account for roughly 1/3 of the peaks detected in a given spectrum but only a small portion of total m/z range. Ringing is a known Fourier transformation artifact and we hypothesize that partial ringing and fuzzy sites are Fourier transformation-based artifacts as well. Partial ringing appears similar to ringing but suppressed at the scan level, with the ringing pattern only visible with multiple scan aggregated. Fuzzy sites have a more distinct pseudo-Gaussian appearance and were only observed in spectra from non-Lumos Fusion Tribid instruments and not in spectra from our SolariX ICR instrument and nor from other Orbitrap instruments.

All three artifacts complicate assignment and confound experimental interpretation; however, our study focused primarily on the fuzzy sites as they differed significantly from known artifacts and were particularly problematic for our classification studies. The correlation between fuzzy site location and sample class increases the probability of class-specific misassignment, introducing sample-specific artifactual features. Ultimately the presence of these artifacts produces brittle classifiers and complicates characterizing the true biological variance between sample classes using mass spectrometry. The results presented in this study reflect both the severity of this artifact for untargeted experiments, but also the deficiency in assignment tools specifically designed for untargeted analyses. The methods and tools presented in this study detect and remove HPD-sites in a non-encoding manner (i.e. will not encode sample class into spectra in the condition that HPD location is class specific), while providing sufficient protection from fuzzy sites with existing assignment methods. Ultimately, better assignment pipelines will be necessary as experiments grow in scale. A spectral analysis approach that can leverage peak correspondence between scans to infer if a peak is artifactual or real would produce higher quality peaklists for downstream data analysis, assignment, and interpretation.

Resource Sharing:

Code and data used for this manuscript are available here: https://figshare.com/s/700ea5fde9c2229c1f9c.

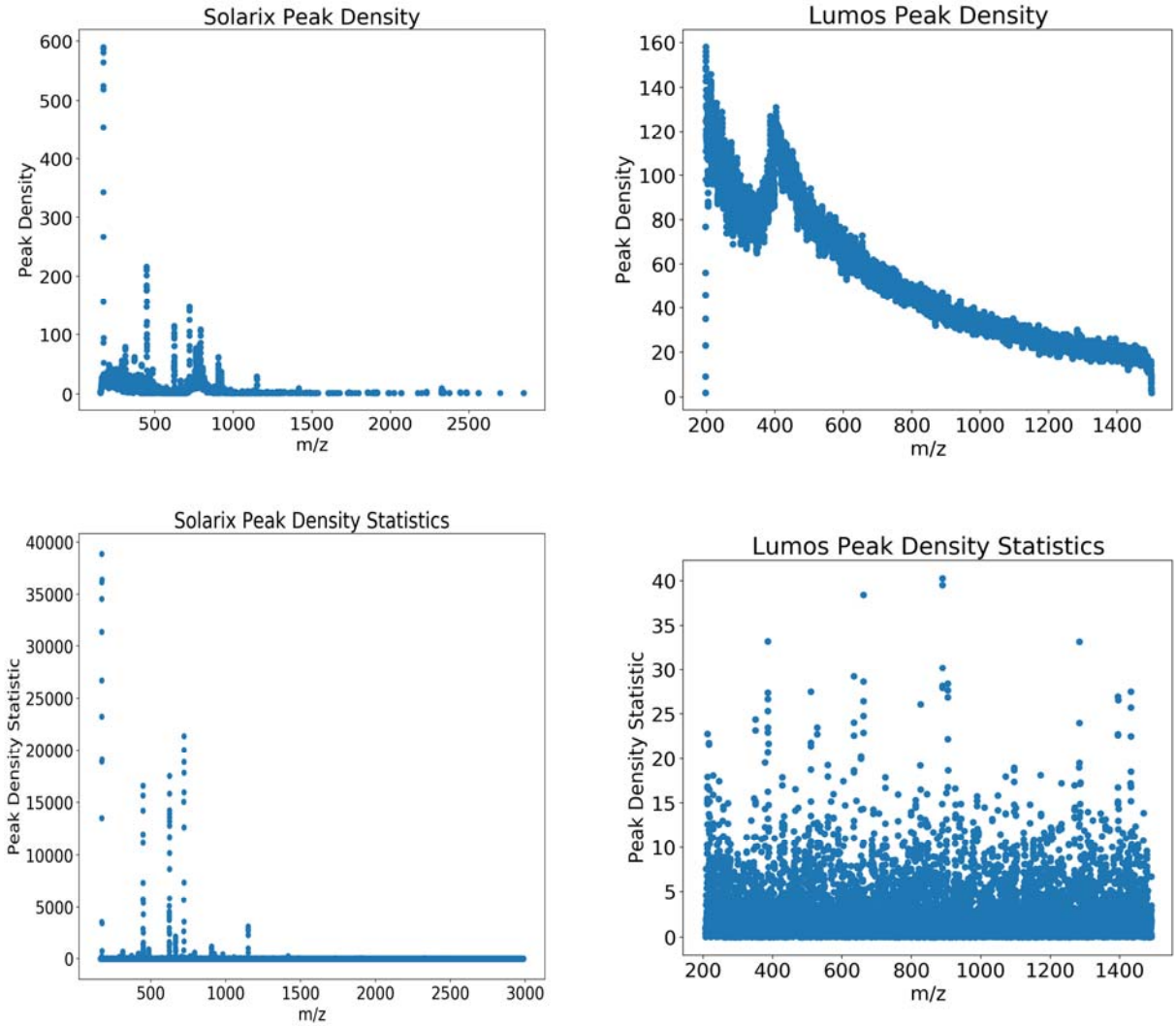
Acknowledgements:

This work was supported in part by National Science Foundation grant NSF 1252893 (Hunter N.B. Moseley), National Institutes of Health grants NIH 1R03CA211835-01 (Chi Wang and Robert Flight), NIH 1R01ES022191-01 (Teresa W.-M. Fan, Richard M. Higashi, Hunter N.B. Moseley, and Michael Nantz) and NIH 1U24DK097215-01A1 (Richard M. Higashi, Teresa W.-M. Fan, Andrew N. Lane, and Hunter N.B. Moseley), and Grant-in-Aid award from the American Heart Association (AHA16GRNT31310020 to Q.J.W.). We would like to thank Wenzhu Zhang and Brian T. Chait in the National Resource for the Mass Spectrometric Analysis of Biological Macromolecule at the Rockefeller University, and Avantika Dhabaria and Beatrix Ueberheide at the Proteomics Resource Center at the University Langone Medical Center for providing us with spectra for our bioinformatics analyses. We would also like to thank Thomas Wilson and the High Resolution Metabolomics Laboratory (HRML) at the Institute of Biological, Environmental and Rural Sciences for Q-Exactive+ data. We would like to thank Timothy Fahrenholz for collection of direct infusion spectra on the Fusion. We would like to thank Qiushi Sun for collection of ICMS spectra on the Fusion. Finally, we would like to thank Mike Senko from ThermoFisher Scientific for the extremely helpful discussions on the known and possible origins of the artifacts we have observed in Orbitrap FTMS instruments.

Supplemental Figure Legends

Supplemental Figure 1: Additional Peak Density Examples

We have performed peak density analyses on both our Solarix ICR instrument and the Thermo Lumos Tribrid Fusion instrument, a more advanced version of the Tribrid Fusion. In the ICR severe HPD phenomena are present due to ringing phenomena. It is not clear if the ringing artifacts from ICR are identical in orgin to those in orbitrap spectra. Our Lumos examples show no obvious HPD artifacts of any kind. In both the ICR and the Lumos, the peak density decreases with increasing peak density but not monotonically.



{
“Contaminants”: [
{
“Formula”: “C18H34O4”,
“Name”: “Dibutyl Sebacate”,
“Mass”: “314.245697”,
“Class”: “Plasticiser”,
},
{
“Formula”: “C8H18O3”,
“Name”: “Dibutyl carbitol”,
“Mass”: “162.125595”,
“Class”: “Scintillation cocktail”,
}, …
]
}

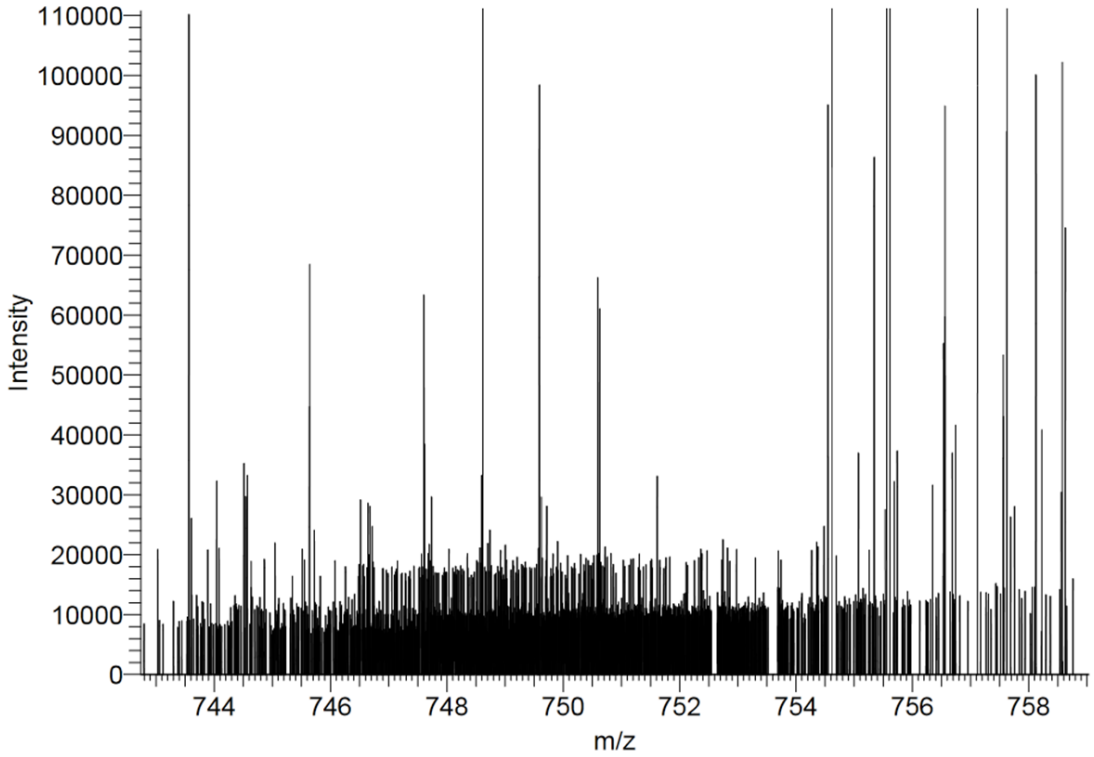
Supplemental Figure 4: Adducts JSON

{
“Components”: {
“[M+1H1-e_POS]”: {
“Max”: 1,
“ExtraGroups”: [“Adducts”],
“Mass”: 1.007276452191,

}

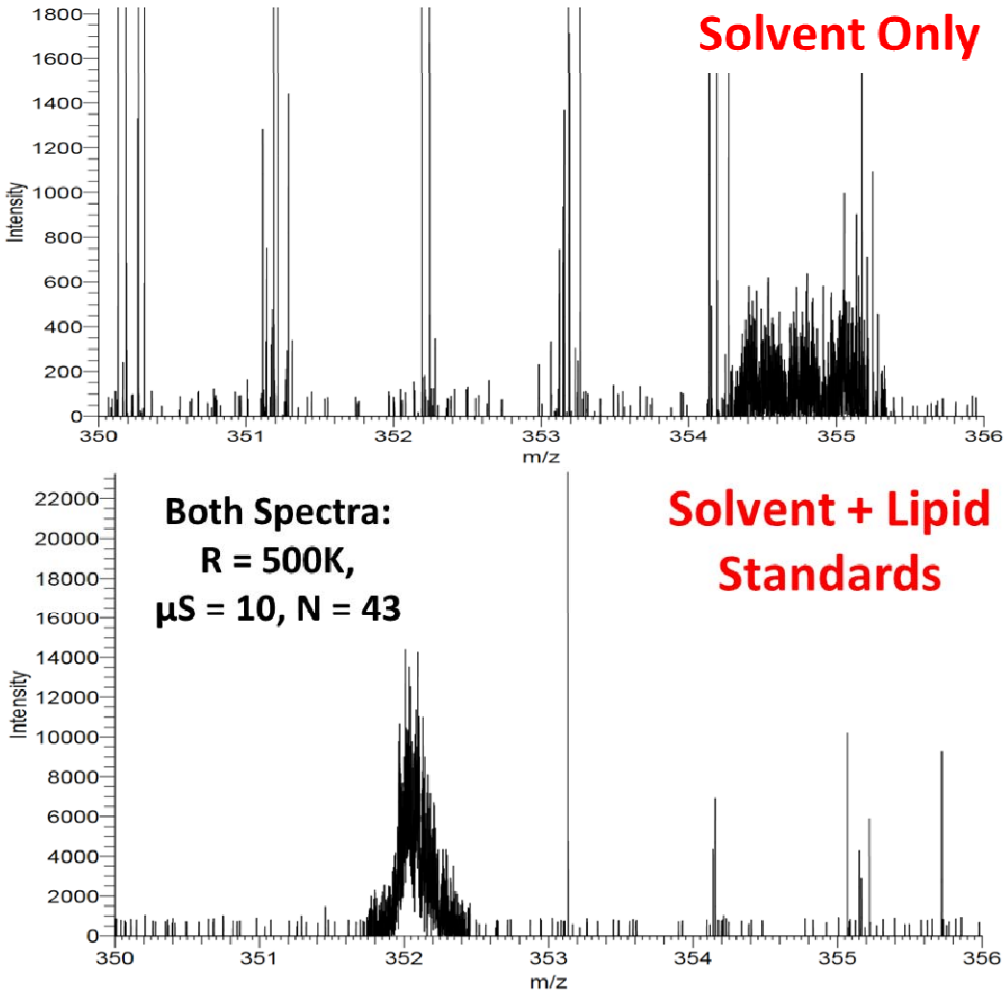
}

Supplemental Figure 5: High uS Fuzzy Site
At very high microscan settings (Sample A), fuzzy sites become almost uniform and span a wide m/z window. Although still responsible for artifactual peaks, the significant intensity difference allows the distinguishing of true signal.



Supplemental Figure 6: FuzzySites vary with Sample Composition

A small change in chemical composition (Sample A with and without lipid standards) changes fuzzy site location. With only solvent, there is a fuzzy site at 354.8 m/z. With lipid standards, this fuzzy site shifts to 352.1 m/z. The number of fuzzy sites will remain constant, but will all be shifted by roughly the same amount.



Supplemental Figure 7: Spectral Acquisition Times for Sample D

Spectra for sample D show no significant “batches” with respect to time. Most samples were run sequentially over a time period from 7/2014 to 9/2015.

Sample #	Class	Start Acquisition Time
1	normal	7/27/2014 10:36
2	cancer	7/27/2014 18:54
3	normal	7/27/2014 19:27
4	normal	7/27/2014 20:35
5	cancer	7/27/2014 21:09
6	normal	7/27/2014 21:43
7	cancer	7/27/2014 22:16
8	normal	7/27/2014 22:50
9	cancer	7/27/2014 23:23
10	normal	7/27/2014 23:57
11	cancer	7/28/2014 0:31
12	normal	7/28/2014 1:05
13	cancer	7/28/2014 1:39
14	cancer	7/28/2014 2:46
15	normal	7/28/2014 3:20
16	cancer	7/28/2014 3:53
17	normal	7/28/2014 4:28
18	cancer	7/28/2014 5:01
19	normal	7/28/2014 5:35
20	cancer	7/28/2014 6:09
21	cancer	7/28/2014 10:39
22	normal	7/28/2014 11:13
23	normal	8/6/2014 13:01
24	normal	8/6/2014 13:46
25	normal	12/19/2014 17:09
26	normal	12/19/2014 18:16
27	cancer	12/19/2014 18:50
28	cancer	12/19/2014 19:58
29	normal	12/19/2014 20:31
30	cancer	12/20/2014 16:26
31	cancer	12/20/2014 17:51
32	cancer	12/20/2014 18:07
33	normal	12/20/2014 18:57
34	cancer	12/20/2014 19:15
35	normal	12/20/2014 19:32
36	cancer	12/20/2014 19:49
37	cancer	1/7/2015 19:38

38	normal	1/7/2015 19:54
39	normal	1/7/2015 21:36
40	normal	1/7/2015 22:09
41	normal	1/7/2015 22:26
42	normal	1/8/2015 12:11
43	cancer	1/8/2015 12:28
44	cancer	1/8/2015 13:19
45	cancer	1/8/2015 13:52
46	normal	1/8/2015 15:37
47	normal	8/18/2015 21:11
48	normal	8/18/2015 21:45
49	cancer	8/18/2015 22:02
50	normal	8/18/2015 22:18
51	cancer	8/18/2015 22:52
52	normal	8/18/2015 23:09
53	cancer	8/18/2015 23:26
54	normal	8/18/2015 23:44
55	cancer	8/19/2015 0:17
56	cancer	8/19/2015 22:49
57	normal	8/19/2015 23:06
58	cancer	8/19/2015 23:22
59	normal	8/19/2015 23:39
60	cancer	8/19/2015 23:56
61	normal	8/20/2015 0:47
62	cancer	8/20/2015 1:04
63	normal	8/20/2015 1:21
64	cancer	8/20/2015 1:38
65	normal	8/20/2015 2:12
66	cancer	8/20/2015 2:28
67	normal	8/20/2015 2:45
68	cancer	8/20/2015 3:02
69	normal	8/20/2015 3:19
70	cancer	8/20/2015 3:36
71	normal	8/20/2015 4:10
72	cancer	8/20/2015 4:27
73	normal	8/20/2015 4:43
74	cancer	8/20/2015 5:00
75	normal	8/20/2015 19:20
76	cancer	8/20/2015 19:37
77	normal	8/20/2015 19:54
78	cancer	8/20/2015 20:10
79	normal	8/20/2015 20:28
80	cancer	8/20/2015 20:44

81	normal	8/20/2015 21:18
82	cancer	8/20/2015 21:34
83	normal	8/20/2015 21:52
84	cancer	8/20/2015 22:08
85	normal	8/20/2015 22:25
86	cancer	8/20/2015 22:42
87	normal	8/20/2015 23:16
88	cancer	8/20/2015 23:33
89	normal	8/20/2015 23:50
90	cancer	8/21/2015 0:06
91	normal	8/21/2015 0:23
92	cancer	8/21/2015 0:41
93	normal	8/21/2015 1:15
94	cancer	8/21/2015 1:31
95	normal	8/21/2015 1:48
96	cancer	8/21/2015 2:05
97	normal	8/21/2015 2:22
98	cancer	8/21/2015 2:38
99	normal	8/21/2015 3:12
100	cancer	8/21/2015 3:30
101	normal	8/21/2015 3:46
102	cancer	8/21/2015 4:03
103	normal	8/21/2015 4:21
104	cancer	8/21/2015 4:37
105	normal	8/21/2015 5:11
106	cancer	8/21/2015 5:28
107	normal	8/22/2015 15:11
108	cancer	8/22/2015 15:28
109	normal	8/22/2015 15:46
110	cancer	8/22/2015 16:03
111	normal	8/22/2015 19:17
112	cancer	8/22/2015 19:34
113	cancer	8/22/2015 20:08
114	normal	9/4/2015 0:14
115	cancer	9/4/2015 0:31
116	normal	9/4/2015 1:05
117	cancer	9/4/2015 1:21
118	normal	9/4/2015 1:39
119	cancer	9/4/2015 1:55
120	normal	9/4/2015 2:12
121	cancer	9/4/2015 2:47
122	normal	9/4/2015 3:04
123	cancer	9/4/2015 3:20

124	normal	9/4/2015 3:37
125	cancer	9/4/2015 3:54
126	normal	9/4/2015 4:27
127	cancer	9/4/2015 4:45
128	normal	9/4/2015 5:02
129	cancer	9/4/2015 5:18
130	normal	9/4/2015 5:36
131	cancer	9/4/2015 6:09
132	normal	9/4/2015 6:26
133	cancer	9/4/2015 6:44
134	normal	9/4/2015 7:00
135	cancer	9/4/2015 7:17
136	normal	9/4/2015 7:50
137	cancer	9/4/2015 8:08
138	normal	9/4/2015 20:14
139	cancer	9/4/2015 20:30
140	normal	9/4/2015 20:48
141	cancer	9/4/2015 21:05
142	normal	9/4/2015 21:21
143	cancer	9/4/2015 21:55
144	normal	9/4/2015 22:12
145	cancer	9/4/2015 22:29
146	normal	9/4/2015 22:46
147	cancer	9/4/2015 23:03
148	normal	9/4/2015 23:36
149	cancer	9/4/2015 23:53
150	normal	9/5/2015 0:10
151	cancer	9/5/2015 0:27
152	normal	9/5/2015 0:44
153	cancer	9/5/2015 1:18
154	normal	9/5/2015 1:34
155	cancer	9/5/2015 1:52
156	normal	9/5/2015 2:08
157	cancer	9/5/2015 2:25
158	normal	9/5/2015 2:59
159	cancer	9/5/2015 3:17
160	normal	9/5/2015 3:33
161	cancer	9/5/2015 3:50
162	normal	9/5/2015 4:07
163	cancer	9/5/2015 4:40
164	normal	9/5/2015 4:58
165	cancer	9/5/2015 5:15
166	normal	9/5/2015 5:31

167	cancer	9/5/2015 5:48
168	normal	9/5/2015 6:22
169	cancer	9/5/2015 6:39
170	normal	9/5/2015 6:56
171	cancer	9/5/2015 7:13
172	normal	9/5/2015 7:30
173	cancer	9/5/2015 8:03
174	normal	9/5/2015 8:20
175	cancer	9/5/2015 8:38
176	normal	9/5/2015 8:54
177	cancer	9/5/2015 9:11
178	normal	9/16/2015 19:22
179	cancer	9/21/2015 20:28
180	normal	9/21/2015 20:45
181	normal	9/23/2015 19:04
182	cancer	9/23/2015 19:22
183	normal	9/23/2015 19:39
184	cancer	9/23/2015 19:56
185	cancer	9/23/2015 20:13
186	cancer	9/23/2015 20:29

References:

Acharjee, A., Prentice, P., Acerini, C., Smith, J., Hughes, I. A., Ong, K., . . . Koulman, A. (2017). The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. *Metabolomics, 13*(3), 25. doi:10.1007/s11306-017-1166-2

Astarita, G., Ahmed, F., & Piomelli, D. (2009). Lipidomic analysis of biological samples by liquid chromatography coupled to mass spectrometry. *Methods Mol Biol, 579*, 201-219. doi:10.1007/978-1-60761-322-0_10

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Carreer, W. J., Flight, R. M., & Moseley, H. N. B. (2013). A Computational Framework for High-Throughput Isotopic Natural Abundance Correction of Omics-Level Ultra-High Resolution FT-MS Datasets. *Metabolites, 3*(4), 853-866. doi:10.3390/metabo3040853

Creek, D. J., Chokkathukalam, A., Jankevics, A., Burgess, K. E., Breitling, R., & Barrett, M. P. (2012). Stable isotope-assisted metabolomics for network-wide metabolic pathway elucidation. *Analytical chemistry, 84*(20), 8442-8447.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). MASS SPECTROMETRY-BASED METABOLOMICS. *Mass spectrometry reviews, 26*(1), 51-78. doi:10.1002/mas.20108

Eyles, S. J., & Kaltashov, I. A. (2004). Methods to study protein dynamics and folding by mass spectrometry. *Methods, 34*(1), 88-99.

Fan, T. W., Lane, A. N., Higashi, R. M., Farag, M. A., Gao, H., Bousamra, M., & Miller, D. M. (2009). Altered regulation of metabolic pathways in human lung cancer discerned by 13C stable isotope-resolved metabolomics (SIRM). *Molecular Cancer, 8*(1), 41. doi:10.1186/1476-4598-8-41

Fan, T. W. M., Lorkiewicz, P., Sellers, K., Moseley, H. N. B., Higashi, R. M., & Lane, A. N. (2012). Stable isotope-resolved metabolomics and applications for drug development. *Pharmacology & therapeutics, 133*(3), 366-391. doi:10.1016/j.pharmthera.2011.12.007

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology, 22*(5), 245-252. doi:http://dx.doi.org/10.1016/j.tibtech.2004.03.007

Griffiths, P. R., & Pariente, G. L. (1986). Introduction to spectral deconvolution. *TrAC Trends in Analytical Chemistry, 5*(8), 209-215.

Guan, S., & Marshall, A. G. (1997). Linear prediction Cholesky decomposition vs Fourier transform spectral analysis for ion cyclotron resonance mass spectrometry. *Analytical chemistry, 69*(6), 1156-1162.

Harris, D. M., Li, L., Chen, M., Lagunero, F. T., Go, V. L. W., & Boros, L. G. (2012). Diverse mechanisms of growth inhibition by luteolin, resveratrol, and quercetin in MIA PaCa-2 cells: a comparative glucose tracer study with the fatty acid synthase inhibitor C75. *Metabolomics, 8*(2), 201-210. doi:10.1007/s11306-011-0300-9

Herres, W., & Gronholz, J. (1984). Understanding FT-IR data processing. *Part, 1*, 352-356.

Higashi, R. M., Fan, T. W.-M., Lorkiewicz, P. K., Moseley, H. N., & Lane, A. N. (2014). Stable isotope-labeled tracers for metabolic pathway elucidation by GC-MS and FT-MS. *Mass Spectrometry in Metabolomics: Methods and Protocols*, 147-167.

Hiller, K., Metallo, C. M., Kelleher, J. K., & Stephanopoulos, G. (2010). Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. *Analytical chemistry, 82*(15), 6621-6628.

Hore, P. (1985). NMR data processing using the maximum entropy method. *Journal of Magnetic Resonance (1969), 62*(3), 561-567.

Hore, P. J. (1985). NMR data processing using the maximum entropy method. *Journal of Magnetic Resonance (1969), 62*(3), 561-567.

Keller, B. O., Sui, J., Young, A. B., & Whittal, R. M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta, 627*(1), 71-81. doi:http://dx.doi.org/10.1016/j.aca.2008.04.043

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*, 234. doi:10.1186/1471-2105-7-234

Lorkiewicz, P., Higashi, R. M., Lane, A. N., & Fan, T. W.-M. (2012). High information throughput analysis of nucleotides and their isotopically enriched isotopologues by direct-infusion FTICR-MS. *Metabolomics, 8*(5), 930-939.

Mathur, R., & O'Connor, P. B. (2009). Artifacts in Fourier transform mass spectrometry. *Rapid communications in mass spectrometry : RCM, 23*(4), 523-529. doi:10.1002/rcm.3904

McLafferty, F. W. (1959). Mass spectrometric analysis. Molecular rearrangements. *Analytical chemistry, 31*(1), 82-87.

Miladinović, S. M., Kozhinov, A. N., Tsybin, O. Y., & Tsybin, Y. O. (2012). Sidebands in Fourier transform ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry, 325-327*, 10-18. doi:10.1016/j.ijms.2012.08.009

Moseley, H. N. (2010). Correcting for the effects of natural abundance in stable isotope resolved metabolomics experiments involving ultra-high resolution mass spectrometry. *BMC Bioinformatics, 11*(1), 139. doi:10.1186/1471-2105-11-139

Moseley, H. N. B. (2013). ERROR ANALYSIS AND PROPAGATION IN METABOLOMICS DATA ANALYSIS. *Computational and Structural Biotechnology Journal, 4*(5), e201301006. doi:https://doi.org/10.5936/csbj.201301006

Ogura, T., Bamba, T., & Fukusaki, E. (2013). Development of a practical metabolite identification technique for non-targeted metabolomics. *Journal of Chromatography A, 1301*, 73-79. doi:http://dx.doi.org/10.1016/j.chroma.2013.05.054

Peake, D. A., Yokoi, Y., Wang, J., & Yingying, H. (2013). A New Lipid Software Workflow for Processing Orbitrap-based Global Lipidomics Data in Translational and Systems Biology Research Retrieved from https://tools.thermofisher.com/content/sfs/posters/New-Lipid-Software-Workflow-PN-ASMS13-Th591-DPeake.pdf

Philip B Tooke, S. F. P. (1989). Resolution, Apodisation And Bandshape In FTIR Spectroscopy. *Proc.SPIE, 1145*, 1145 - 1145 - 1142. doi:10.1117/12.969637

Ren, J.-G., Seth, P., Clish, C. B., Lorkiewicz, P. K., Higashi, R. M., Lane, A. N., . . . Sukhatme, V. P. (2014). Knockdown of malic enzyme 2 suppresses lung tumor growth, induces differentiation and impacts PI3K/AKT signaling. *Scientific reports, 4*.

Sellers, K., Fox, M. P., Michael Bousamra, II, Slone, S. P., Higashi, R. M., Miller, D. M., . . . Deshpande, R. (2015a). Pyruvate carboxylase is critical for non–small-cell lung cancer proliferation. *The Journal of clinical investigation, 125*(2), 687.

Sellers, K., Fox, M. P., Michael Bousamra, I., Slone, S. P., Higashi, R. M., Miller, D. M., . . . Deshpande, R. (2015b). Pyruvate carboxylase is critical for non–small-cell lung cancer proliferation. *The Journal of clinical investigation, 125*(2), 687.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*: Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Walt, S. v. d., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering, 13*(2), 22-30.

Weber, R. J., Li, E., Bruty, J., He, S., & Viant, M. R. (2012). MaConDa: a publicly accessible mass spectrometry contaminants database. *Bioinformatics, 28*(21), 2856-2857.

Wood, M. L., & Mark Henkelman, R. (1985). Truncation artifacts in magnetic resonance imaging. *Magnetic resonance in medicine, 2*(6), 517-526.

Yang, Y., Fan, T. W. M., Lane, A. N., & Higashi, R. M. (2017a). Chloroformate derivatization for tracing the fate of Amino acids in cells and tissues by multiple stable isotope resolved metabolomics (mSIRM). *Analytica Chimica Acta, 976*, 63-73. doi:http://dx.doi.org/10.1016/j.aca.2017.04.014

Yang, Y., Fan, T. W. M., Lane, A. N., & Higashi, R. M. (2017b). Chloroformate derivatization for tracing the fate of Amino acids in cells and tissues by multiple stable isotope resolved metabolomics (mSIRM). *Analytica Chimica Acta*(976), 63-73.

Zhang, J., & Brodbelt, J. S. (2004). Screening flavonoid metabolites of naringin and narirutin in urine after human consumption of grapefruit juice by LC-MS and LC-MS/MS. *Analyst, 129*(12), 1227-1233. doi:10.1039/B412577K