LA-UR-23-25282

Approved for public release; distribution is unlimited.

Title: Fast Semi-automated Filtration Method for Non-targeted LC-QTOF Data of Aged Nitroplasticizer Samples

Author(s): Chen, Kitmin
Yang, Dali
Edgar, Alexander Steven

Intended for: Propellants, Explosives, Pyrotechnics

Issued: 2023-07-25 (Rev.1) (Draft)

F u l l P a p e r

Fast Semi-Automated Filtration Method for Non-Targeted LC-QTOF Data of Aged Nitroplasticizer Samples

Kitmin Chen[a], Alexander S. Edgar[a], Dali Yang[a]*

Abstract: A full dataset of aged nitroplasticizer (NP) is composed of more than 2000 unique mass-to-charges (m/z) when combining the non-targeted data obtained from both positive and negative electrospray ionization modes in time-of-flight mass spectrometry. Therefore, manual processing of these data often takes days, weeks, or even months to scrutinize for mechanistic insights. To effectively extract meaningful signals that represent vital degradation intermediates in the early NP degradation mechanism, a semi-automated post-processing workflow for data filtering, tailored to the aging experiment of NP, has been developed. The automated portion of this workflow is written in a Python code (using pandas, numpy, and matplotlib libraries), which removes more than 65% of potential false signals within seconds via four threshold-based adjustable filters: signal sensitivity, coefficient of variation, number of measurements, and retention time variability. As for the manual portion, a pattern-based inspection method is employed to reduce another 23% or more false positives, which greatly simplifies data visualization and results in less than 3% of potential candidate m/z needing in-depth data interpretation. As a positive control, known compounds are verified. Using this semi-automated data reduction method, the amount of time required is reduced to a matter of hours for data filtering in the non-targeted datasets of aged NP, which saves more time and effort for compound identification.

Keywords: Nitroplasticizer (NP), Aging, Time-of-flight mass spectrometry (TOF-MS), Data automation, Python

1 Introduction

Nitroplasticizer (NP) is a binder component commonly used in energetic composites to reduce mechanical sensitivity. The NP studied here is a eutectic mixture of bis(2,2-dinitropropyl) acetal/formal (BDNPA/F) (~1:1 wt. ratio), which also contains approximately 0.1 wt.% of *n*-Phenyl-*β*-naphthylamine (PBNA) for scavenging NO_x oxidant species generated from NP decomposition. In a 44-month thermal aging study of NP [1-4], the investigation of decomposition products is crucial to understand the life-time aging behavior of NP when it is in decades-long storage. To detect these unspecified compounds, liquid chromatography quadrupole time-of-flight mass spectrometry (LC-QTOF) has been applied by using information-dependent acquisition (IDA), which performs a non-targeted survey scan of precursor ions and selects the candidate ions based on the preset criteria (e.g., intensity threshold) for secondary fragmentation (via collision-induced dissociation) [5-8]. Subsequently, the recorded mass-to-charges ratios (m/z) of precursor and fragment ions in MS^1 and MS^2 (also called MS/MS) spectra are used to identify the detected compounds. Through the non-targeted MS data, major breakthroughs have been achieved in our previous works, including the identification of 2,2-dinitropropanol (DNPOH) [9,10], nitrated derivatives of PBNA and nitrophenols [11,12], and with the potential to carry out quantitation [12]. Leveraging the results of these previous studies, the depletion of dinitro-PBNA derivative is the transition point where NP degradation alters from nitrous acid (HONO) elimination into acid-catalyzed acetal/formal hydrolysis [11,13-17]. Considering other species may be involved in this alteration of the NP degradation pathway, our attention turns to the early stage of NP aging (i.e., the time before NP hydrolysis occurs).

Manual data processing and annotation are very time-consuming because thousands of m/z are typically collected in non-targeted data acquisition [18-20]. Besides the detected m/z of interest, which includes their corresponding isotopes, different adducts, charge states, and in-source fragments, a portion of m/z can be originated

[a] *K. Chen, A.S. Edgar, D. Yang*
MST-7: Engineered Materials Group, Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, United States
**E-mail: dyang@lanl.gov*

from false positives, such as contaminants (in sample, solvent, autosampler vials, etc.) and ionization-generated artefacts [18,20,21]. To eliminate the false positives, various filter-based automated tools have been developed for non-targeted metabolomics studies. For instance, the filtering method published by McMillan et al. used mass defect and retention time (T_R) alignment to remove salt cluster artefacts in metabolomics data [20]. Another filtering approach, the comprehensive peak characterization (CPC) algorithm was presented by Pirttila et al., which also removes noise artefacts and other false peaks based on the peak quality (e.g., signal-to-noise, peak width, and peak intensity) [18]. However, these algorithms are designed for metabolomics applications and their implementation for our purpose may require deep knowledge in coding. Therefore, an automated method tailored to the aging study of NP is needed. Unlike "omics" studies, the aging study of NP monitors the changes in chemical composition on a time scale. Therefore, the changes in abundance as a function of time provide a unique characteristic to investigate the outstanding decomposed products in the aging experiment. As an example, the patterns of rise and fall in the abundance of nitrated PBNA derivatives across 44 months of aging at different temperatures (38, 45, 55, and 64°C) provide insights into how PBNA is consumed in the aged NP samples in our work [11]. As opposed to non-targeted "omics" studies, the identification of a few but vital analytes is a faster alternative than profiling all species. In addition, the identified few signals can be used to expand the search of other products by association. For example, nitrophenol and dinitrophenol were discovered based on the assumption that acid-catalyzed hydrolysis has occurred in the nitrated PBNA derivatives [11]. Additionally, the nitrated PBNA derivatives and nitrophenols were detected despite their abundance combined being less than 0.1 wt.% in NP in the chemical composition, which suggests the vital decomposed products are comparatively higher in abundance as NP decomposes over time. To apply the findings from previous works, a data filtration method using user-specified parameters is preferred. Thus, we can obtain a potential list of vital decomposed products in the early stage of NP aging, where each represents a unique m/z and retention time value and exhibits a meaningful correlation between its abundance and aging time. To achieve this objective, we explored the use of the Python programming language. In this paper, we present a fast semi-automated data filtration method for the aging study of NP and a demonstrative evaluation of the real non-targeted datasets obtained from large sets of aged NP samples.

2 Experimental Section

2.1 Evaluation of Data Quality

Before testing the Python script on the dataset, an evaluation of the aged NP data in both electrospray ionization modes (ESI +/−) was conducted through cross-examination of known compounds between the new and old methods. The optimization of the LC-QTOF parameters and the spectral quality (i.e., T_R and mass accuracy) in ESI− are detailed in previous works [11,12]. By standardizing the sample preparation method [11] and instrumental calibration [12], the drawn curves (patterns) of the relative intensities are smoother in the new dataset (top row) than the old dataset (bottom row). Therefore, the new dataset can provide better data mining capabilities to find anomalies and correlations.

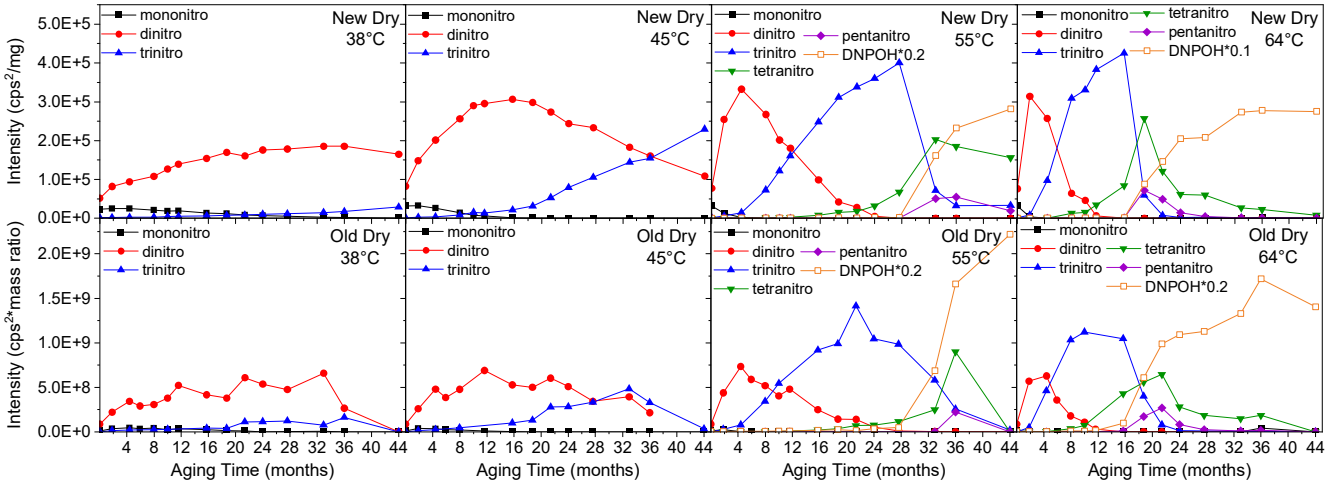


Figure 1. Using the relative intensities of DNPOH and PBNA nitro derivatives, pattern comparison is conducted between new (top row) and old (bottom row) ESI− datasets in dry-aged NP from 0 to 44 months.

Though the previous improvements of ESI+ provided confirmation of a mononitro-PBNA derivative [11], the sensitivity for this compound is still 4-fold lower than its detection in ESI−. Therefore, most of the discoveries were conducted in ESI− [9,11]. However, this also means the ESI+ data may contain important mechanistic insights that are not covered by the ESI− data (e.g., PBNA, despite being a known stabilizer in NP, it was not detected until ESI+ mode is employed [11]). Therefore, the LC parameters for ESI+ acquisition was further optimized, as described in **Table 1**. With the starting composition of mobile phases changed from 30% B to 60% B (where B = 0.1% formic acid in methanol) and the flow rate decreased from 0.35 mL/min to 0.30 mL/min, the intensity in ESI+ has significantly improved: 98-fold in PBNA, 22-fold in mononitro-PBNA (3 times the ESI− signal), 50-fold in dinitro-PBNA (2 times the ESI− signal), and 45-fold in trinitro-PBNA. The overall data quality is crucial to the semi-automated filtration analysis not only in the aspect of enhanced sensitivity, but also in the aspect of T_R and mass defect, which maintained at less than 0.1 min and 5 ppm (<10 ppm in fragments).

Table 1. LC Parameters (ExionLC AC)

LC method components	Old LC parameter in ESI+ mode	New LC parameters in ESI+ mode	LC parameters in ESI− mode
Mobile phases	A: 0.1% formic acid in water B: 0.1% formic acid in methanol	A: 0.1% formic acid in water B: 0.1% formic acid in methanol	A: 13 mM ammonium acetate in water, pH 6.0 B: 13 mM ammonium acetate in 95:5 (v/v) ACN, pH 6.0
Needle rinse	0.02% formic acid and 5% acetone in ACN	0.02% formic acid and 5% acetone in ACN	0.02% formic acid and 5% acetone in ACN
Run time (min)	20	16	18
Gradient program (%B, curve[a])	0.00 min (30.0%, -1), 10.0 min (99.9%)	0.00 min (60.0%, -1), 10.0 min (99.9%), 10.1 min (99.9%), 12.6 min (99.9%), 12.7 min (60.0%)	0.00 min (20.0%), 3.00 min (60.0%), 10.0 min (60.0%), 10.1 min (99.9%), 14.0 min (99.9%), 14.1 min (60.0%)
Column rinse and equilibration program	0.00 min (30% B), 0.01 min (99.9%B), 4.05 min (99.9% B), 4.10 min (30% B), 10.0 min (30% B)	Combined with gradient program as a single method	Combined with gradient program as a single method
Flow rate (mL/min)	Acquisition and equilibration = 0.35 Column rinse = 0.50	Acquisition and equilibration = 0.30 Column rinse = 0.50	Acquisition and equilibration = 0.35 Column rinse = 0.53

[a] If not specified, the curve is set to 0 or linear.

2.2 Data Processing

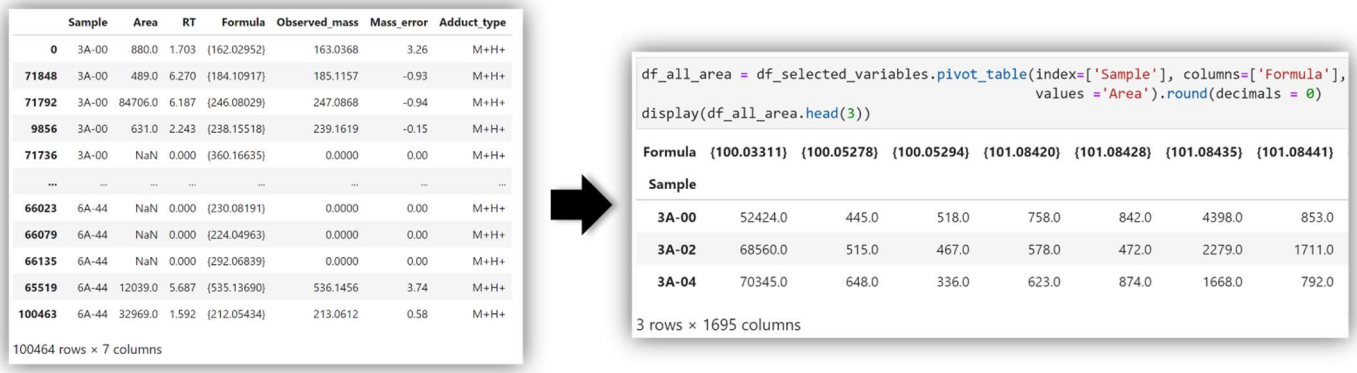
The processing procedure of MS data is divided into four components: peak finding, peak integration, data reduction, and data compilation. Using the peak finding and integration features of the default SCIEX OS software, four testing datasets (ESI+/−) are obtained from aged NP in wet and dry environments. Non-targeted peak finding was performed at mid-exhaustive sensitivity to capture all plausible m/z between 2.0 and 7.5 min in ESI− and 1.0 and 8.5 min in ESI+. To minimize integration of baseline noises, the parameters of the selected m/z in the SCIEX OS software were modified: the minimum peak height at 200 counts per second (cps), the signal-to-noise threshold at 10, the gaussian smoothing at 2.5, the baseline noise at 40%, the baseline subtraction window at 1 min, and the peak splitting at 4 points. The semi-automated filtration method is composed of three steps: (1) using the first Python script, the anomalies are automatically removed by four adjustable filters, including the sensitivity of peak intensity, the relative changes of intensities across the aging process, the minimal amount of measurements, and the T_R variability; (2) from the filtered list of m/z, the relative intensities of each m/z are plotted against the sample names (i.e., approximated aging time) to allow visual inspection to be conducted; and (3) using the second Python script, the spectrometric information of the accepted m/z is automatically compiled as an Excel summary.

3 Results and Discussion

Combining both the ESI+ and the ESI− datasets, manual cleaning of more than 2000 m/z is prone to introducing human error. As shown in the example in **Figure 2**, the ESI+ dataset alone contain 100464 rows of information and there are 1695 unique m/z after data extraction. Using pandas, numpy, and matplotlib, the automated post-filtration process offers a fast, simple, and targeted approach for data reduction, which involves four different filters in a stepwise progression. It is worth noting that the initial number of m/z can be greatly

reduced with strict integration parameters and narrow time windows of peak screening, which also avoids counting impurities in the dead volume and column rinse. However, the peak finding and integration settings in the SCIEX OS software were intentionally relaxed to obtain the maximum number of m/z, and thus the limitation of the Python script can be tested.

As mentioned previously, our primary objective is to search for the vital decomposed products in the early stage of NP aging. Based on the knowledge gained from the low-abundant compounds (i.e., nitrated PBNA derivatives and nitrophenols), the first three filters define what is considered vital, that is, Filter 1, the measured maximum abundance (intensity) of decomposition product must be comparable to the lowest detected value obtained from the low-abundant compounds if not higher; Filter 2, its abundance must change in response to NP degradation; and Filter 3, the change in abundance is expected to occur as a function of time, the rate of which can be potentially measured. In the last filter, T_R alignment is used to minimize random error. Consequently, the order of filters is arranged following this logic. However, because Filter 3 cannot determine whether the change in abundance is consecutive or random, visual inspection is applied.



	Sample	Area	RT	Formula	Observed_mass	Mass_error	Adduct_type
0	3A-00	880.0	1.703	{162.02952}	163.0368	3.26	M+H+
71848	3A-00	489.0	6.270	{184.10917}	185.1157	-0.93	M+H+
71792	3A-00	84706.0	6.187	{246.08029}	247.0868	-0.94	M+H+
9856	3A-00	631.0	2.243	{238.15518}	239.1619	-0.15	M+H+
71736	3A-00	NaN	0.000	{360.16635}	0.0000	0.00	M+H+
...	...	...	...	...	...	...	...
66023	6A-44	NaN	0.000	{230.08191}	0.0000	0.00	M+H+
66079	6A-44	NaN	0.000	{224.04963}	0.0000	0.00	M+H+
66135	6A-44	NaN	0.000	{292.06839}	0.0000	0.00	M+H+
65519	6A-44	12039.0	5.687	{535.13690}	536.1456	3.74	M+H+
100463	6A-44	32969.0	1.592	{212.05434}	213.0612	0.58	M+H+

100464 rows × 7 columns

```
df_all_area = df_selected_variables.pivot_table(index=['Sample'], columns=['Formula'],
                                                values ='Area').round(decimals = 0)
display(df_all_area.head(3))
```

Formula	{100.03311}	{100.05278}	{100.05294}	{101.08420}	{101.08428}	{101.08435}	{101.08441}
Sample							
3A-00	52424.0	445.0	518.0	758.0	842.0	4398.0	853.0
3A-02	68560.0	515.0	467.0	578.0	472.0	2279.0	1711.0
3A-04	70345.0	648.0	336.0	623.0	874.0	1668.0	792.0

3 rows × 1695 columns

Figure 2. Snapshots of the data filtration workflow in JupyterLab. The raw non-targeted data of dry-aged NP are exported from SCIEX OS as a comma separated value (csv) file, which contains 100464 rows of data (left). Using the unique formulas (m/z + adduct) and the samples names of 44-month thermal aging experiment of NP as the parameters for columns and rows, respectively, their corresponding intensity values are extracted into a table (right) that has a dimension of 56 rows and 1695 columns. The first three rows are displayed as an example.

3.1 Filter 1 – Sensitivity

The SCIEX OS software will integrate the baseline noises if an inadequate threshold is used. In addition, a minimum intensity in IDA-based method is required to perform MS^2 scans, which produces the fragmentation pattern that can be used for structural elucidation to increase identification confidence. When only an MS^1 spectrum is available, formulaic estimation is still achievable and the isotopic pattern can be used for confirmation of elemental composition. Ideally, obtaining the MS^2 is preferred because the exact structure of the decomposed product is critical to trace its origin. Therefore, the threshold setting in Filter 1 was calibrated using the lowest detected intensity for generating an MS^2 spectrum of the least sensitive known compound (i.e., trinitro-PBNA derivative in ESI+ and nitrophenol in ESI−). Using the improved ESI+ parameters, the number of m/z found in ESI+ is at least twice the number of m/z found in ESI−, as shown in **Table 2**. Therefore, the intensity threshold was set 3.5-fold higher in ESI+ data (35000 cps^2) than in ESI− data (10000 cps^2), as shown in **Figure 3** (i.e., filter1_intensity in Python script), which removed more than 64% of ESI+ data and more than 34% of ESI− data that are false positives.

3.2 Filter 2 – Percentage of Variation

Filter 2 discerns and removes the impurities and contaminants that could be mistakenly treated as valid m/z. As previously mentioned, similar to the nitrated PBNA derivatives in the thermal aging process, the concentrations of intermediates are expected to change (increase or decrease) as a function of time in aging process. For species that exist in the baseline NP sample and do not break down or do not break down significantly throughout 44 months of aging, especially at elevated temperatures (>55°C), they are identified either as impurities or production byproducts that are very persistent. For species that are detected at approximately the same intensity in the solvent blank (ACN), the baseline NP sample, and all aged NP samples, they are identified as contaminants that likely originated from external sources, such as solvent, mobile phases, autosampler vials, etc. Using the coefficient of variation (CV) expressed in percentage (i.e., standard deviation ÷ mean × 100%), the second filter determines whether the intensity changes are statistically significant. Although the

CV of mono, di, and trinitro-PBNA derivatives are measured at 82%, 63%, and 146% in ESI− and 77%, 50%, and 112% in ESI+ across 56 measurements of dry-aged NP (excluded solvent blanks and standards), the CV of impurities typically fall between 10% and 40%. However, the CV threshold must also avoid false negatives that could contribute to the alteration of NP degradation, particularly intermediates that can be produced by minor reactions at low temperatures. Therefore, the change in intensity of nitrophenol across 44 months of aging at 38°C (i.e., 13 measurements) is examined because it is proposed as the minor hydrolyzed product of nitrated PBNA derivatives [11], which yields a CV of 16%. Based on these results, a fixed 15% was implemented as the CV threshold, as shown in **Figure 3** (i.e., filter2_impurity in Python script). Consequently, about 2% of false positives in ESI+ data and 8% or more of false positives in ESI− data are removed.

3.3 Filter 3 – Number of Measurements

Filter 3 minimizes the random species, including artefacts (e.g., ionization-generated artefacts) and species that do not contribute to the NP degradation pathway. In species that do contribute to the NP degradation pathway, their abundance typically changes as a function of time in a consecutive manner and their rate of change follows the time temperature superposition (TTS) principle. Therefore, multiple consecutive measurements can be obtained (e.g., the rise and fall of nitrated PBNA derivatives and the formation rate of DNPOH as a key product of NP hydrolysis that are accelerated by elevated temperatures [11]). Meanwhile, some species only occur at random occasions and therefore do not contribute to the NP degradation pathway. To generate any pattern of relative changes in intensity at various temperatures, a minimum of eight measurements is required (i.e., two measurements per temperature). Using this estimated value as the threshold, which is shown in **Figure 3** (i.e., filter3_minimal_measurements in Python script), 1%-2% and 2%-7% false positives are removed in ESI+ and ESI− datasets, respectively. Considering more than 20 measurements in mono, di, and trinitro derivatives are obtained across 56 aged NP samples in both ESI+ and ESI−, the minimum tolerance can be adjusted to a higher number for further data reduction if needed.

3.4 Filter 4 – T_R Variability

Because the noise and artefacts are randomized defects that tend to have varying T_R, the fourth filter is a complementary layer to all three filters described above. Filter 4 removes any randomized defects that exhibits high variability in T_R. Through meticulous control of the external factors (e.g., temperature, column equilibration time, and mobile phases) that could affect T_R reproducibility, low T_R variability of less than 3 s or 0.05 min was achieved in the datasets of a previous study [11]. Because of a faster change of the mobile phase composition in the gradient program of the ESI− method (>7% B per min) than that of the ESI+ method (>4% B per min), stronger T_R drift was observed in the early eluents of ESI− dataset, such as 0.032 min in DNPOH as opposed to 0.011 min in BDNPA/F [11]. Therefore, the cut-off for T_R variability of Filter 4 was set to a slightly wider window (5 s) for the ESI− dataset whereas 3 s was selected for the ESI+ dataset, as shown in **Figure 3** (i.e., filter4_RT_drift in Python script), which removes about 11%-16% of ESI− data and 5%-7% of ESI+ data that are false positives.

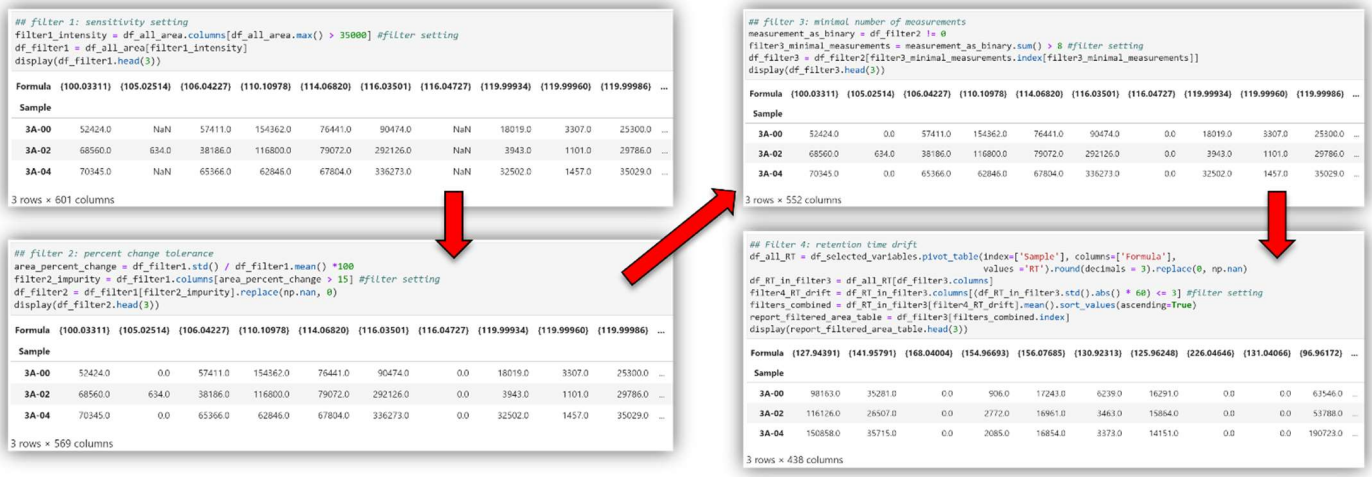


Figure 3. Snapshots of the automated data filtering method in JupyterLab. Through the four filters, data are reduced from 1695 columns to 438 columns, where each represents a unique m/z. The first three rows in the results of each filter are displayed as an example.

Table 2. Semi-automated data filtration results

	m/z Quantity in dry-aged NP datasets		m/z Quantity in wet-aged NP datasets	
Stages of semi-automated filtration	**ESI+**	**ESI−**	**ESI+**	**ESI−**
Initial number of m/z	1695	655	1635	775
After Filter 1	601	428	551	437
After Filter 2	569	351	524	374
After Filter 3	552	308	496	361
After Filter 4	438	205	408	272
Total candidate m/z	11	9	33	17

Through all four filters, the automated process reduced the initial number of m/z by more than 74% in the ESI+ datasets and 65% in the ESI− datasets. Because the objective of this process is to search for vital decomposed products that are high in abundance, the majority of the false positives must be removed by Filter 1. Using conservative settings, the contributions of Filters 2 and 3 to the removal of false positives are merely 3%-4% in ESI+ dataset. When the settings are doubled in value (30% CV and 16 points of minimal measurements), their contributions are increased to 14% (6% by Filter 2 and 8% by Filter 3) and the number of filtered m/z drop from 438 to 277: a significant difference of 161 m/z that does not require manual inspection.

3.5 Visual Inspection of Abundance Change in m/z

The visual inspection of abundance change in m/z is a pattern-based qualitative validation. The plots are displayed as a top-down list (ESI− signals first, then ESI+ signals) in the order of T_R that matches the Excel checklist (generated by Python script). The validation is governed by three principles and examples of false signals are illustrated in **Figures 4** and **5**: (1) any peaks with a pattern resembling baseline noise, system contaminants, and sample impurities are removed; (2) any peaks with a pattern that only occurs in the quadrants of elevated temperatures (≥ 55°C, after 4ACN) are known as high-temperature intermediates and are therefore removed since they do not contribute to the early aging of NP; and (3) isotopes and in-source fragments are removed (or archived on a separate list), which exhibit the same pattern and T_R as the monoisotopic mass. Following these basic criteria, more than 23% and 30% false positives in ESI+ and ESI− datasets are removed, respectively.

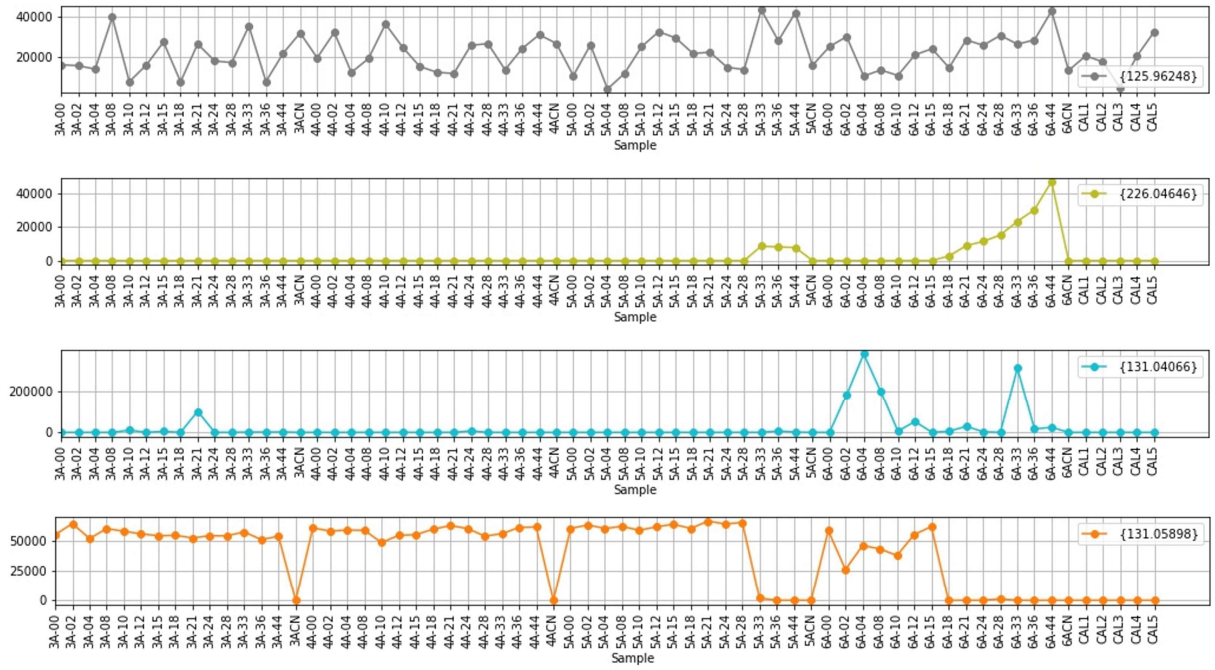


Figure 4. Plots of peak intensity (y-axis) versus the aged samples in the chronological order of aging months (x-axis) that is given by the last two digits (e.g., 02 = 2 months). These are example plots of baseline noise (first row), high-temperature intermediates (second row), random defects (third row), and impurity (fourth row). Each plot is divided into four quadrants by the ACN blanks (i.e., 3ACN, 4ACN, 5ACN, and 6ACN) from left to right, with each quadrant representing a temperature that is given by the first number of the sample name (i.e., 3 = 38°C, 4 = 45°C, 5 = 55°C, and 6 = 64°C). The number displayed in the legend is the estimated mass of the chemical formula and not a m/z value.

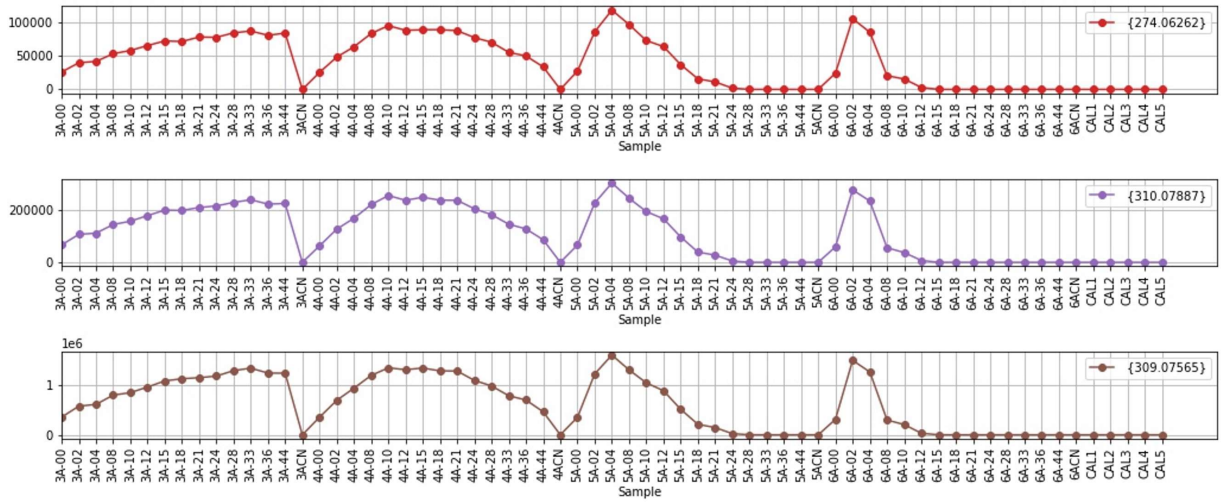


Figure 5. The plots of dinitro-PBNA derivative (bottom), its isotope (middle), and in-source fragment (top).

After the visual inspection process, a targeted search was performed in the SCIEX OS software to confirm the validity of m/z and retrieve the MS^1 and MS^2 spectra. Although the chemical structure is a crucial property to understand the reaction mechanism, the interpretation of MS^2 features is a challenging and time-consuming process, especially for chimeric spectra, which contain false fragment features from the precursors of similar m/z that pass through the mass filter [22]. Therefore, only the chemical formulas with unspecified adducts are determined for the candidate m/z, as presented in **Tables 3 and 4**. In **Table 5**, 100% of the previously identified intermediates [9,11] are captured in the filtered lists. Beside using these known compounds as calibrants to optimize the filtration parameters for maximum effectiveness, which is discussed in detail above, they can also be used as a positive confirmation.

Table 3. Filtered list of candidate m/z from ESI+ datasets for future structural interpretation

T_R (min)	Observed m/z in dry-aged NP (Da)	Observed m/z in wet-aged NP (Da)	Estimated formula	Observation
2.080	ND	159.0442	$C_{10}H_7O_2^+$	
2.182	273.0698	ND	$C_8H_{11}N_5O_6^+$	Chimeric MS^1/MS^2; other possibility: $C_8H_{14}N_2O_7Na^+$
2.631	ND	285.1045	$C_7H_{17}N_4O_8^+$	$C_7H_{13}N_3O_8$, $[M + NH_4]^+$
2.631	ND	290.0600	$C_7H_{13}N_3O_8Na^+$	$C_7H_{13}N_3O_8$, $[M + Na]^+$
2.643	ND	329.1306	$C_9H_{21}N_4O_9^+$	$C_9H_{17}N_3O_9$, $[M + NH_4]^+$
2.643	ND	334.0864	$C_9H_{17}N_3O_9Na^+$	$C_9H_{17}N_3O_9$, $[M + Na]^+$
2.996	ND	343.1101	$C_9H_{19}N_4O_{10}^+$	$C_9H_{15}N_3O_{10}$, $[M + NH_4]^+$
2.996	ND	348.0654	$C_9H_{15}N_3O_{10}Na^+$	$C_9H_{15}N_3O_{10}$, $[M + Na]^+$
3.015	ND	256.1164	$C_{10}H_{16}N_4O_4^+$	
3.052	ND	214.0504	$C_{12}H_8NO_3^+$	Chimeric MS^1/MS^2
3.087	ND	149.0922	$C_5H_{13}N_2O_3^+$	Chimeric MS^1/MS^2
3.177	276.1192	276.1195	$C_{10}H_{18}N_3O_6^+$	
3.181	ND	287.0857	$C_9H_{13}N_5O_6^+$	
3.422	ND	268.1144	$C_8H_{18}N_3O_7^+$	
3.450	228.1961	ND	$C_{13}H_{26}NO_2^+$	Chimeric MS^1/MS^2
3.735	ND	331.0987	$C_{10}H_{16}N_4O_4^+$	
3.929	250.0868	250.0863	$C_{16}H_{12}NO_2^+$	
3.995	ND	225.1964	$C_{13}H_{25}N_2O^+$	
4.322	ND	247.0865	$C_{16}H_{11}N_2O^+$	
4.376	ND	248.1074	$C_{17}H_{14}NO^+$	
4.717	ND	393.0875	$C_{10}H_{15}N_7O_{10}^+$	Chimeric MS^1/MS^2
4.719	ND	489.1428	$C_{13}H_{25}N_6O_{14}^+$	Other possibility: $C_{14}H_{24}N_7O_{11}Na^+$
5.111	ND	503.1590	$C_{15}H_{26}N_7O_{11}Na^+$	Other possibility: $C_{14}H_{27}N_6O_{14}^+$
5.188	219.1043	219.1041	$C_{16}H_{13}N^+$	
5.242	249.1027	249.1028	$C_{16}H_{13}N_2O^+$	
5.286	ND	249.1083	$C_9H_{17}N_2O_6^+$	Other possibility: $C_{10}H_{16}N_3O_3Na^+$
5.878	ND	474.3064	$C_{22}H_{42}N_4O_7^+$	Other possibility: $C_{25}H_{43}N_2O_5Na^+$

6.466	276.0772	ND	$C_{16}H_{10}N_3O_2^+$	Other possibility: $C_{18}H_{12}O_3^+$
6.659	ND	274.1588	$C_{20}H_{20}N^+$	
6.665	ND	233.1935	undetermined	Low-intensity chimeric MS^1/MS^2
7.051	ND	274.1593	$C_{20}H_{20}N^+$	
ND = Not detected				

Table 4. Filtered list of candidate m/z from ESI− datasets for future structural interpretation

T_R (min)	Observed m/z in dry-aged NP (Da)	Observed m/z in wet-aged NP (Da)	Estimated formula
3.961	ND	233.0208	$C_{10}H_5N_2O_5^-$
4.477	ND	182.0216	$C_6H_4N_3O_4^-$
5.104	ND	244.0772	$C_{17}H_{10}NO^-$
5.114	ND	198.0197	$C_{11}H_4NO_3^-$
5.172	ND	370.1097	$C_9H_{18}N_6O_{10}^-$
5.323	ND	349.0576	$C_{17}H_9N_4O_5^-$
5.391	ND	258.0514	$C_{12}H_8N_3O_4^-$
5.392	289.0618	289.0618	$C_{17}H_9N_2O_3^-$
5.535	ND	384.0897	$C_{12}H_{14}N_7O_8^-$
5.546	ND	222.0526	$C_9H_8N_3O_4^-$

Table 5. Verification of known compounds found in the filtered list

T_R (min)	ESI mode	Estimated formula, adduct type	Observed m/z in dry-aged NP (Da)	Observed m/z in wet-aged NP (Da)	Compound
2.010	−	$C_6H_4N_2O_5$, $[M-H]^-$	183.0047	183.0048	dinitrophenol
3.283	−	$C_3H_6N_2O_5$, $[M-CH_3O]^-$	119.0101	119.0098	DNPOH (as in-source fragment)
3.974	−	$C_6H_5NO_3$, $[M-H]^-$	ND	138.0198	nitrophenol
5.557	−	$C_{16}H_8N_6O_{10}$, $[M-H]^-$	443.0220	ND	pentanitro-PBNA
6.069	−	$C_{16}H_9N_5O_8$, $[M-H]^-$	398.0364	398.0381	tetranitro-PBNA
6.102	−	$C_{16}H_{12}N_2O_2$, $[M-H]^-$	263.0817	263.0824	mononitro-PBNA
6.106	−	$C_{16}H_{11}N_3O_4$, $[M-H]^-$	308.0667	308.0677	dinitro-PBNA
6.212	−	$C_{16}H_{10}N_4O_6$, $[M-H]^-$	353.0521	353.0522	trinitro-PBNA
4.923	+	$C_{16}H_{11}N_3O_4$, $[M+H]^+$	310.0825	ND	dinitro-PBNA
4.909	+	$C_{16}H_{11}N_3O_4$, $[M]^+$	ND	309.0751	dinitro-PBNA
4.912	+	$C_{16}H_{11}N_3O_4$, $[M+Na]^+$	ND	332.0647	dinitro-PBNA
5.173	+	$C_{16}H_{10}N_4O_6$, $[M+H]^+$	355.0687	355.0683	trinitro-PBNA
5.248	+	$C_{16}H_{12}N_2O_2$, $[M+H]^+$	265.1016	265.0976	mononitro-PBNA
5.391	+	$C_{16}H_{13}N$, $[M+H]^+$	220.0967	220.1035	PBNA

4 Conclusion

With less than 80 lines of Python code, the automation of data filtering is achieved. By applying simple filters and a pattern-based visual inspection, the semi-automated data filtration method simplifies the review process of non-targeted data obtained from the aged NP samples, which allows a faster transition into the compound identification stage. To extract spectrometric information of vital degradation intermediates that are mildly sensitive at low temperatures, the developed workflow effectively removes up to 97% of the false positives among the more than 2000 m/z detected in both ESI+ and ESI− modes. Based on the verification result of the previous identified compounds [9,11,12], the capability and performance of the Python code is demonstrated. Moreover, the four filters described above are adjustable and therefore can be further optimized when we gain deeper knowledge of NP degradation or can be calibrated to suit specific study needs. However, this developed method may not apply to other non-targeted studies because it is specifically tailored to study the datasets collected from aged NP. Future focus should be on the expansion of data automation via Python, such as adding more filters to account for other variables (e.g., mass accuracy) and replacing the manual visual inspection process with mathematic algorithms (e.g., linear, quadratic, exponential, power, etc.) to evaluate the changes in relative intensities.

Acknowledgements

We thank Justine Yang, Jillian O'Neel, Joseph Torres, and Camille Wong for their experimental work preceding this publication. We also thank David Langlois for the synthesis of DNPOH, and Kevin Morris (Consolidated Nuclear Security Pantex) for his parallel work of NP evaluation. This work was supported by the US

Department of Energy through the Los Alamos National Laboratory Aging and Lifetimes Program. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

References

[1] D. Yang, A.S. Edgar, J.A. Torres, J.C. Adams, J.D. Kress, "Thermal Stability of a Eutectic Mixture of Bis(2,2-dinitropropyl) Acetal and Formal: Part C. Kinetic Compensation Effect," *Propellants Explos. Pyrotech.* **2020,** *46* (1), 134-149.
[2] D. Yang, R. Pacheco, S. Edwards, K. Henderson, R. Wu, A. Labouriau, P. Stark, "Thermal stability of a eutectic mixture of bis(2,2-dinitropropyl) acetal and formal: Part A. Degradation mechanisms in air and under nitrogen atmosphere," *Polym. Degrad. Stab.* **2016,** *129*, 380-398.
[3] D. Yang, R. Pacheco, S. Edwards, J. Torres, K. Henderson, M. Sykora, P. Stark, S. Larson, "Thermal stability of a eutectic mixture of bis(2,2-dinitropropyl) acetal and formal: Part B. Degradation mechanisms under water and high humidity environments," *Polym. Degrad. Stab.* **2016,** *130*, 338-347.
[4] D. Yang, D.Z. Zhang, "Role of water in degradation of nitroplasticizer," *Polym. Degrad. Stab.* **2019,** *170*.
[5] X. Zhu, Y. Chen, R. Subramanian, "Comparison of information-dependent acquisition, SWATH, and MS(All) techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry," *Anal. Chem.* **2014,** *86* (2), 1202-9.
[6] J.D. Whitman, K.L. Lynch, "Optimization and Comparison of Information-Dependent Acquisition (IDA) to Sequential Window Acquisition of All Theoretical Fragment Ion Spectra (SWATH) for High-Resolution Mass Spectrometry in Clinical Toxicology," *Clin Chem* **2019,** *65* (7), 862-870.
[7] G. de Albuquerque Cavalcanti, R. Moreira Borges, G. Reis Alves Carneiro, M. Costa Padilha, H.M. Gualberto Pereira, "Variable Data Independent Acquisition and Data Mining Exploring Feature-Based Molecular Networking Analysis for Untargeted Screening of Synthetic Cannabinoids in Oral Fluid," *J. Am. Soc. Mass Spectrom.* **2021,** *32* (9), 2417-2424.
[8] T.N. Decaestecker, S.R. Vande Casteele, P.E. Wallemacq, C.H. Van Peteghem, D.L. Defore, J.F. Van Bocxlaer, "Information-Dependent Acquisition-Mediated LC−MS/MS Screening Procedure with Semiquantitative Potential," *Analytical Chemistry* **2004,** *76* (21), 6365-6373.
[9] A.S. Edgar, C.H. Wong, K. Chen, D.A. Langlois, D. Yang, "Identification of 2,2-dinitropropanol, a Hydrolyzed Product of Aged Eutectic Bis(2,2-dinitropropyl) Acetal – Bis(2,2-dinitropropyl) Formal Mixture," *Propellants Explos. Pyrotech.* **2022**.
[10] C.H. Wong, A.S. Edgar, D. Yang, "Liquid Chromatography Mass Spectrometry Study of a Eutectic Mixture of bis(2,2-Dinitropropyl) Acetal/Formal," *Propellants Explos. Pyrotech.* **2021,** *46* (12), 1849-1859.
[11] K. Chen, A.S. Edgar, J. Jung, J.D. Kress, C.H. Wong, D. Yang, "Liquid Chromatography Quadrupole Time-of-Flight Mass Spectrometry Analysis of Eutectic Bis(2,2-dinitropropyl) Acetal/Formal Degradation Profile: Nontargeted Identification of Antioxidant Derivatives," *ACS Omega* **2022,** *7* (39), 35316-35325.
[12] K. Chen, A.S. Edgar, C.H. Wong, D. Yang, "Liquid Chromatography Quadrupole Time-of-Flight Mass Spectrometry: A Strategy for Optimization, Characterization, and Quantification of Antioxidant Nitro Derivatives," *ACS Omega* **2022,** *7* (36), 32701-32707.
[13] J. Clayden, N. Greeves, S. Warren, "Nucleophilic Substitution at C=O with Loss of Carbonyl Oxygen," in *Organic Chemistry*, 2nd ed. (Oxford University Press, New York, 2012), Chap. 11, pp 223-228.
[14] J. Clayden, N. Greeves, S. Warren, "Determining Reaction Mechanisms," in *Organic Chemistry*, 2nd ed. (Oxford University Press, New York, 2012), Chap. 39, pp 1058-1059.
[15] C.E. Freye, C.J. Snyder, "Investigation into the Decomposition Pathways of an Acetal-Based Plasticizer," *ACS Omega* **2022,** *7* (34), 30275-30280.
[16] J.D. Kress, Nitroplasticizer Resistance to Hydrolysis *(Personal Communication)*. Los Alamos National Laboratory, Los Alamos, NM, United States, 2006.
[17] P.W. Leonard, Mechanism of acetal (or formal) hydrolysis *(Personal Communication)*. Los Alamos National Laboratory, Los Alamos, NM, United States, 2016.
[18] K. Pirttila, D. Balgoma, J. Rainer, C. Pettersson, M. Hedeland, C. Brunius, "Comprehensive Peak Characterization (CPC) in Untargeted LC-MS Analysis," *Metabolites* **2022,** *12* (2).
[19] A. Z. Woldegiorgis, "LC-MS/MS Based Metabolomics to Identify Biomarkers Unique to Laetiporus sulphureus," *International Journal of Nutrition and Food Sciences* **2015,** *4* (2).
[20] A. McMillan, J.B. Renaud, G.B. Gloor, G. Reid, M.W. Sumarah, "Post-acquisition filtering of salt cluster artefacts for LC-MS based human metabolomic studies," *J Cheminform* **2016,** *8* (1), 44.
[21] L. Vaclavik, O. Lacina, J. Hajslova, J. Zweigenbaum, "The use of high performance liquid chromatography-quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety," *Anal Chim Acta* **2011,** *685* (1), 45-51.
[22] S. Xing, H. Yu, M. Liu, Q. Jia, Z. Sun, M. Fang, T. Huan, "Recognizing Contamination Fragment Ions in Liquid Chromatography-Tandem Mass Spectrometry Data," *J. Am. Soc. Mass Spectrom.* **2021,** *32* (9), 2296-2305.

F u l l P a p e r

Supporting Information

Fast Semi-Automated Filtration Method for Non-Targeted LC-QTOF Data of Aged Nitroplasticizer Samples

Kitmin Chen[a], Alexander S. Edgar[a], Dali Yang[a]*

MST-7: Engineered Materials Group, Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States.

Python codes developed for this work; run by JupyterLab 3.0.14:

```
## Semi-automated m/z filtration
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

## Load csv file and reformat data presentation
filename = "C:\\Users\\...dataset.csv"
raw_data = pd.read_csv(filename, sep=',', header=1, index_col=0).reset_index()
raw_data['Sample Name'] = raw_data['Sample Name'].str[4:]
raw_data['Retention Time'] = raw_data['Retention Time'].str.replace('N/A', '0')
raw_data['Retention Time'] = pd.to_numeric(raw_data['Retention Time'])
raw_data['Found At Mass'] = raw_data['Found At Mass'].str.replace('N/A', '0')
raw_data['Found At Mass'] = pd.to_numeric(raw_data['Found At Mass'])
raw_data['Mass Error'] = raw_data['Mass Error'].str.replace('N/A', '0')
raw_data['Mass Error'] = pd.to_numeric(raw_data['Mass Error'])
raw_data = raw_data.set_index('Sample Name')
#raw_data = raw_data.drop(index = ['ACN', 'Blank', 'CAL'], axis = 1) #Samples can be removed if needed
raw_data = raw_data.reset_index().sort_values('Sample Name')

df_selected_variables = raw_data[["Sample Name", "Area", "Retention Time", "Formula", "Found At Mass", "Mass Error",
                                  "Adduct/Charge"]].rename(columns = {"Sample Name":"Sample", "Retention Time":"RT",
                                  "Found At Mass":"Observed_mass", "Mass Error":"Mass_error",
                                  "Adduct/Charge":"Adduct_type"})
## filter 1: sensitivity setting
df_all_area = df_selected_variables.pivot_table(index=['Sample'], columns=['Formula'], values ='Area').round(decimals = 0)
filter1_intensity = df_all_area.columns[df_all_area.max() > 35000] #filter setting (cps^2): 35000 for ESI+ and 10000 for ESI−
df_filter1 = df_all_area[filter1_intensity]

## filter 2: percent change tolerance
area_percent_change = df_filter1.std() / df_filter1.mean() *100
filter2_impurity = df_filter1.columns[area_percent_change > 15] #filter setting (%)
df_filter2 = df_filter1[filter2_impurity].replace(np.nan, 0)

## filter 3: minimal number of measurements
measurement_as_binary = df_filter2 != 0
filter3_minimal_measurements = measurement_as_binary.sum() > 8 #filter setting
df_filter3 = df_filter2[filter3_minimal_measurements.index[filter3_minimal_measurements]]

## Filter 4: retention time drift
df_all_RT = df_selected_variables.pivot_table(index=['Sample'], columns=['Formula'], values ='RT').round(decimals = 3).replace(0, np.nan)
df_RT_in_filter3 = df_all_RT[df_filter3.columns]
filter4_RT_drift = df_RT_in_filter3.columns[(df_RT_in_filter3.std().abs() * 60) <= 5] #filter setting (seconds)
filters_combined = df_RT_in_filter3[filter4_RT_drift].mean().sort_values(ascending=True)

# ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
report_filtered_area_table = df_filter3[filters_combined.index] #for report use
report_filtered_RT_table = df_RT_in_filter3[report_filtered_area_table.columns] #for report use
# ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## data compilation as report summary
component1_RT = filters_combined.reset_index().rename(columns = {0:"avg_RT"})
```

Full Paper

```
source_adduct_list = df_selected_variables[["Formula", "Adduct_type"]].set_index('Formula')
component2_adducts = source_adduct_list.loc[filters_combined.index].reset_index()
df_merged_components = pd.merge(component1_RT, component2_adducts, on = 'Formula', how = 'left').drop_duplicates()
df_merged_components['avg_RT'] = df_merged_components['avg_RT'].round(decimals = 3)
source_mass_list = df_selected_variables.pivot_table(index=['Sample'], columns=['Formula'], values ='Observed_mass')
# ------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
report_filtered_mass_table = source_mass_list[filters_combined.index].replace(0, np.nan) #for report use
#  ------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
component3_masses = report_filtered_mass_table.mean().round(decimals=4).reset_index().rename(columns = {0:"avg_mass"})
report_data_summary = pd.merge(df_merged_components, component3_masses, on = 'Formula', how = 'left').drop_duplicates() #for report use
#  ------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## data export
with pd.ExcelWriter("Excel_Checklist.xlsx") as excel_file:
      report_data_summary.to_excel(excel_file, "Checklist for visual inspection")
      report_filtered_area_table.to_excel(excel_file, "peak area")
      report_filtered_mass_table.to_excel(excel_file, "observed mass")
      report_filtered_RT_table.to_excel(excel_file, "observed RT")

## generate plots for visual inspection
fig_height = min(len(report_filtered_area_table.columns) * 2.5, 900)
fig_width = len(report_filtered_area_table.index) / 4.5
xtick_max = len(report_filtered_area_table.index)
report_filtered_area_table.plot(grid = True, marker='o', figsize = (fig_width, fig_height),
                                subplots = True, sharex = False, xlim = 0,
                                xticks = np.arange(0, xtick_max, 1), rot = 90)
plt.tight_layout()
plt.savefig('visual inspection.jpeg')
print("Report/Checklist is now available!")


## Load final candidates for data retrieval and summary
filename2 = "C:\\Users\\...Selected_import.csv"
accepted_data = pd.read_csv(filename2, sep=',', header=1, index_col='Formula')

finalized_area_list = report_filtered_area_table[accepted_data.index]
finalized_mass_list = report_filtered_mass_table[accepted_data.index]
finalized_RT_list = report_filtered_RT_table[accepted_data.index]

with pd.ExcelWriter("Finalized_POS_List.xlsx") as excel_file:
          accepted_data.to_excel(excel_file, "summary")
          finalized_area_list.to_excel(excel_file, "peak area")
          finalized_mass_list.to_excel(excel_file, "observed mass")
          finalized_RT_list.to_excel(excel_file, "observed RT")

## generate final plots
finalized_masses = finalized_area_list.reset_index()
fig_height = min(len(report_filtered_area_table.columns) * 2.5, 900)
fig_width = len(report_filtered_area_table.index) / 4.5
xtick_max = len(report_filtered_area_table.index)
finalized_masses.plot(grid = True, marker='o', figsize = (fig_width, fig_height),
                                subplots = True, sharex = False, xlim = 0,
                                xticks = np.arange(0, xtick_max, 1), rot = 90)
plt.tight_layout()
plt.savefig('final plot.jpeg')
```