© American Society for Mass Spectrometry, 2018

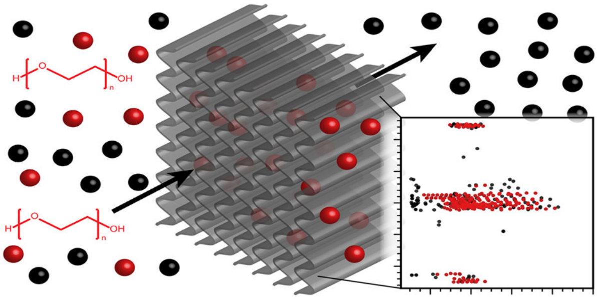
J. Am. Soc. Mass Spectrom. (2019) 30:268–277
DOI: 10.1007/s13361-018-2069-9

RESEARCH ARTICLE

Computational Removal of Undesired Mass Spectral Features Possessing Repeat Units via a Kendrick Mass Filter

Ricardo R. da Silva,[1,2] Fernando Vargas,[1,2] Madeleine Ernst,[1,2] Ngoc Hung Nguyen,[1,2] Sanjana Bolleddu,[1,2] Krizia Karen del Rosario,[1] Shirley M. Tsunoda,[1] Pieter C. Dorrestein,[1,2] Alan K. Jarmusch[1,2]

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
[2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0751, USA



Abstract. Polymers are a common component of chemical background which complicates data analysis and can impair interpretation. Undesired chemical background cannot always be addressed via pre-analytical methods, chromatography, or existing data processing methods. The Kendrick mass filter (KMF) is presented for the computational removal of undesired signals present in MS^1 spectra. The KMF is analogous to mass defect filtering but utilizes homology information via Kendrick mass scaling in combination with chromatographic retention time and the number of observed signals. The KMF is intended to assist in situations in which current data processing methods to remove background, e.g., blank subtraction, are either not possible or effective. The major parameters affecting KMF were investigated using PEG 400 and NIST standard reference material 1950 (metabolites in human plasma). Further exploration of the KMF performance was tested using an extract of a swab known to contain polymers. An illustrative real-world example of skin analysis with polymeric signal is discussed. The KMF is also able to provide a high-level view of the compositionality of data regarding the presence of signals with repeat units and indicate the presence of different polymers.

Keywords: Mass spectrometry, Data processing, Polymer science

Received: 14 July 2018/Revised: 29 August 2018/Accepted: 10 September 2018/Published Online: 28 September 2018

Introduction

Mass spectrometry (MS) studies are prone to undesired chemical background. One source of undesired chemical background is polymers, such as polyethylene glycol (PEG). Undesired polymer background can often be avoided by a trained scientist under controlled laboratory conditions; however, the task of avoiding undesired polymer background is more difficult when collecting samples outside of the laboratory (e.g., sample collection by citizen-scientists). The typical mass spectrum that results from the presence of polymers is complex and consists of oligomer signals separated by the mass-to-charge (*m/z*) of the polymer unit repeat, viz., PEG spectra will contain ions corresponding to $C_{2n}H_{4n+2}O_{n+1}$. The presence of polymers and similar undesired chemical background can be so impactful as to preclude data interpretation; therefore, methods to remove such interferences are needed. Pre-analytical methods, e.g., solid phase extraction, are a common means by which to negate the effect of unwanted background; however, such methods can be costly, time consuming, and often modify the molecular composition of the sample. Another possibility is to remove interfering chemical background through chromatographic methods. Compensation for chemical background can also be performed by data processing

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13361-018-2069-9) contains supplementary material, which is available to authorized users.

Correspondence to: Alan Jarmusch; *e-mail:* ajarmusch@ucsd.edu

methods, e.g., blank subtraction; however, this approach relies on co-analyzing samples which faithfully recapitulate the source of the undesired chemical background. When the source is unknown or not anticipated in the experimental design, removal of undesired chemical background is challenging. Here, we propose using the Kendrick mass filter (KMF) to assist in computationally removing undesired polymer signals. The KMF is intended to address the following gap: (i) the polymer background cannot be removed by pre-analytical methods (or modification of the molecular composition is unwanted, e.g. untargeted metabolomics); (ii) data processing via blank subtraction is not possible as the source cannot be faithfully recapitulated; or (iii) cases in which data has already been collected but rendered useless by undesired chemical background.

The Kendrick mass filter combines Kendrick mass scaling with mass defect filtering (MDF). [1] The Kendrick mass is calculated by rescaling the *m/z* of each ion to an integer value of the unit repeat, differing from the IUPAC definition (i.e., ^{12}C is equal to 12 unified atomic mass units). The defect, i.e., Kendrick mass defect (KMD), between the Kendrick scaled *m/z* and the integer Kendrick mass value (i.e., rounded Kendrick scaled *m/z*) is similar between homologous compounds. Homologous compounds can be readily identified by plotting the integer Kendrick values versus the KMD, providing an interpretable scatterplot, in which homologous compounds are horizontally aligned. Kendrick mass plots and similar visualizations, e.g., Van Krevelen diagrams, have been applied in the fields of petroleomics [2], dissolved organic matter [3], and other complex mixtures. Improvement to the visualization of Kendrick mass plot continues, most recently with the introduction of fractional base units which improves the visual resolving power of polymers. [4] Mass defect filtering (MDF) has been used to perform selection and removal of data centered around a user-defined mass defect, calculated using the IUPAC mass scale [5]. MDF has been applied in the study of drug metabolism [6], removal of salt clusters in LC-MS metabolomics data [7], and natural product chemistry [8]. The KMF is rooted in MDF analysis but utilizes additional information via the KMD that can be used to determine homology. Here, we report the proof-of-concept for the computational removal of undesired mass spectral features possessing repeat units by the KMF.

Experimental

NIST standard reference material 1950 metabolites in frozen human plasma, [9] polyethylene glycol 400 (PEG 400), swab extracts, and human skin samples collected using swabs were analyzed using liquid chromatography–mass spectrometry using a quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific) or quadrupole time-of-flight (ToF) mass spectrometer (maXis Impact, Bruker). Sample preparation and instrumental parameters are detailed in the Supplementary Information. QExactive files (.raw) were converted to .mzXML via MSConvert [10]. The qToF files (.d) were exported using DataAnalysis (Bruker) as .mzXML files after lock mass correction. MS^1 feature finding was performed subsequently in MZmine2 [11], with parameters described in the Supplementary Information. The MS^1 feature matrix for NIST plasma, swab extract, PEG 400, swab spiked PEG 400, and swab spiked plasma was split and individual matrices were compared. Shared variables as well as variables for which all peak area values were zero were excluded. The same operations were performed on data collected at 17,500 and 70,000 mass resolution on the QExactive.

Kendrick mass filtering is performed by importing the MS^1 feature matrix and rescaling the *m/z* values according to the unit repeat (Eq. S1), viz., when filtering for PEG, the measured *m/z* are scaled by 44 divided by 44.0262. The rescaled data is then used to calculate the KMD via subtraction of the exact Kendrick mass from the nominal Kendrick mass (Eq. S2). All possible Kendrick scaled differences in *m/z* are calculated pairwise. The difference between two peaks must be the integer repeat unit (e.g., PEG, 44) and the ΔKMD less than the user-defined ΔKMD. The retention time (RT) window (e.g., 60 s) is used in determining filtering. In addition to ΔKMD and retention time exclusion criteria, the pairwise comparison of MS features, which meet ΔKMD and RT criteria is used to determine the number of observed signals (NOS), Fig. S1. The final matrix is exported as comma-separated values (.csv) file.

The Kendrick mass filter (KMF) computationally removes ions (or MS^1 features if performing chemical separation prior to MS), regardless of spectral abundance, from mass spectral data, which meet filtering criteria. Our implementation of the KMF allows the user to customize three parameters: Kendrick mass defect (ΔKMD), chromatographic retention time (RT), and the number of observed signals (NOS). We tested all combinations of the following parameter values: ΔKMD, 0.001, 0.0015, 0.002, 0.0025, 0.005, 0.0067, 0.0075, 0.01, 0.0125, 0.015, 0.02, 0.025, 0.033, 0.05, 0.067, 0.1; RT, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00, 1.25, 1.50, 1.75, 2.00; and NOS, 2, 3, 4, 5, 6. Scaling of mass spectral data to the Kendrick mass scale and computation of the KMD provides a common visual output, the Kendrick mass plot, in which homology is represented by horizontal alignment of points. Kendrick mass scaling, or any rescaling, preserves the *m/z* differences present in the samples. The rationale for scaling in the KMF algorithm, beyond the visual benefit of the Kendrick mass plots, is that a ΔKMD value can be defined rather than defining a slope and y-axis offset as is necessary if the non-scaled mass spectral data is used, illustrated by the Kendrick mass plot and mass defect plot of PEG 400 in Fig. S2. Retention time is only relevant when chemical separation is performed prior to mass spectrometric analysis. The rationale behind inclusion of the retention time is that the probability of non-desired filtering is likely to increase as the RT criterion increases. In the case of polymers, retention time is influenced by the oligomer length and chemical composition. Fundamentally, each oligomer can be separated in time using chromatography; however, this is impractical and tangential when analyzing samples for non-polymeric molecules. The inclusion of the number of observed signals (NOS) criterion is intended to increase the stringency of filtering. Only spectral peaks (or MS^1 features) which have a sufficient number, defined by the user, of observed peaks (or MS^1 features) with $\Delta m/z$ equal to the repeat unit will meet the criterion. The minimum NOS value is 2 which corresponds to a pair of oligomer signals. The rationale is that if homologous molecules with repeat units are present, then multiple oligomer ions will be detected as is the case in mass spectra of polymers. Fundamentally, the greater the NOS equates to more specific filtering; the cost of specificity is the potential for features to remain unfiltered (loss of sensitivity). In practice, the balance between non-specific filtering versus specific removal of undesired signals (MS^1 features) must be assessed on a case-by-case basis.

The KMF and associated plots are available on GitHub (https://github.com/DorresteinLaboratory/Kendrick_Mass_Filter). The KMF was written in R and is available as a Jupyter notebook. Data discussed in this manuscript are publically available at MassIVE (http://massive.ucsd.edu) via the following MassIVE IDs: MSV000081544 and MSV000081548. Principal component analysis (PCA) was performed in R using the pcaMethods package applying the Nonlinear Iterative Partial Least Squares algorithm, after Pareto scaling [12]. Figures were generated with R standard plot function and ggplot2 package and formatted in Adobe Illustrator.

Results and Discussion

Optimization for PEG 400 Removal While Minimizing Feature Removal from NIST Plasma Standard Reference Material

The effect of KMF parameters will vary with different data and parameters should be selected carefully, balancing removal of features with specificity. We defined the following goal: maximize the filtering of PEG 400 features while minimizing the number of plasma features filtered. We evaluated parameter selection using plasma spiked with PEG 400. The ideal parameters in this instance were determined by plotting the ratio of PEG 400 to plasma MS^1 features filtered versus the number of PEG 400 MS^1 features filtered; the points were colored by the MS^1 features filtered from the plasma spiked with PEG 400 (Fig. 1). Triplicate technical measurements of PEG 400, plasma, and plasma spiked with PEG 400 were used. Each point in the plot represents a different set of KMF parameters (restricted to parameters tested). The uppermost grouping corresponded to the greatest filtering of PEG 400 features (y-axis) as well as the

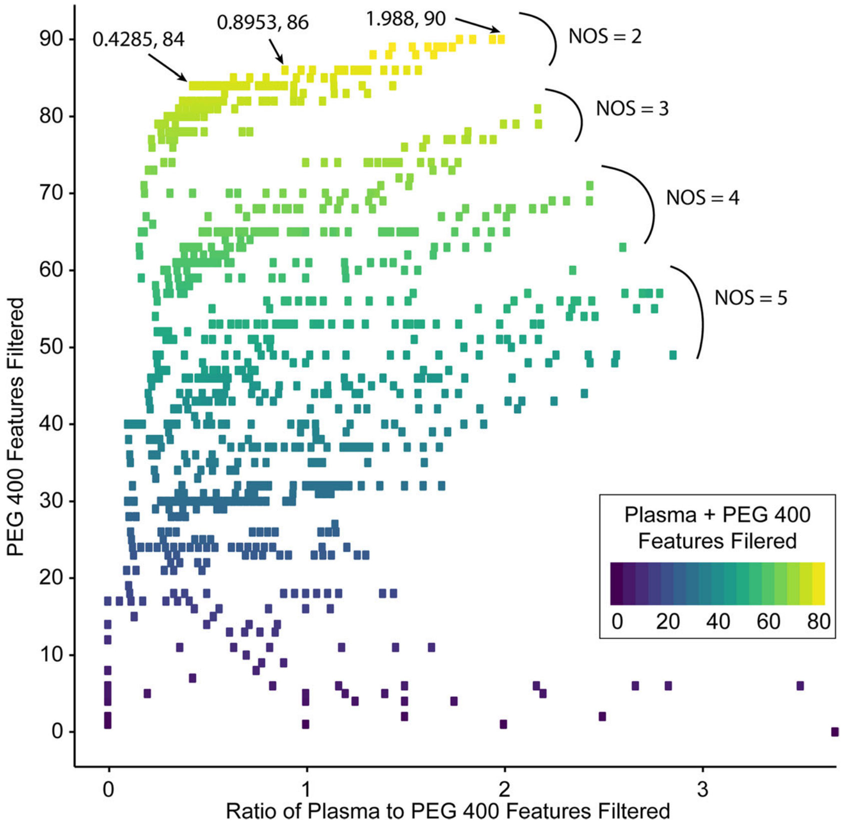


Figure 1. Plot displaying KMF results for all sets, represented as points, of KMF parameters (ΔKMD, RT, and NOS) tested. Data were collected at a resolution of 17,500. KMF parameters for the points highlighted are as follows: 0.4285, 84 (ΔKMD of 0.01, RT of 0.8 min, and NOS of 2); 0.8953, 86 (ΔKMD of 0.033, RT of 0.8 min, and NOS of 2); 1.988, 90 (ΔKMD of 0.1, RT of 2.0 min, and NOS of 2)

greatest number of filtered features in the plasma spiked with PEG 400 sample (colored light-green and yellow). The vertical groupings result from different NOS values. In fulfilling the defined goal (the smallest ratio), the leftmost points in the uppermost grouping are the most suitable sets of KMF parameters. Contrastingly, if filtering is desired with no regard to potential over filtering, then points (sets of KMF parameters) in the rightmost and uppermost grouping and their respective parameters should be used. An interactive plot was created (see Experimental), which displayed the x and y values when hovering over a point. The values were used to look up the associated set of KMF parameters (ΔKMD, RT, and NOS) in the optimization table (electronic supplementary information—Table S1) and displayed in Fig. S3. The parameters chosen (associated with the point 0.4285, 84 in Fig. 1) were as follows: ΔKMD of 0.01, RT of 0.8 min, and NOS of 2.

The effect of KMF parameters (ΔKMD, RT, and NOS) on the number of MS^1 features filtered was systematically explored using PEG 400 and NIST 1950 plasma standard reference material (SRM) based on the previously determined KMF parameters. The effect of ΔKMD is displayed in Fig. 2a-c, isolating RT as a variable (held constant at 0.8 min). PEG 400 MS^1 features filtered, i.e., those meeting the KMF criteria (Fig. 2a), were substantially affected by the ΔKMD parameter. The number of features filtered quickly increased and arrived at a plateau resulting from the removal of all apparent PEG features. The filtering of plasma MS^1 features, not desired, was less affected by ΔKMD values < 0.015, Fig. 2b, but filtering increased with increasing ΔKMD. The NOS parameter, increasing from 2 (red) to 6 (purple), reduced the overall number of features filtered in PEG 400 as well as plasma which reflects the intended increase in stringency. Figure 2c, the overall number of MS^1 features filtered were less with a NOS of 6 compared to a NOS of 2 (most to least stringent, respectively), but a similar ratio of PEG 400 to plasma features is obtained at the local minimum (~ 0.01) which mirrors the ΔKMD parameter chosen during optimization.

The effect of RT on filtering is presented in Fig. 2e-f, isolating ΔKMD (value held constant at 0.01). The filtering of PEG 400 MS^1 features increased with a larger RT parameter. The RT parameter is expected to change based on the chromatographic separation. Figure 2d, filtering increased quickly with small RT parameter (< 0.25 min) increases, afterwards filtering of features plateau with a retention time greater than 0.5 min. The filtering of plasma MS^1 features, Fig. 2e, increased more slowly at small RT parameter values comparatively. This observation supports that PEG is selectively filtered, but the probability of selecting non-

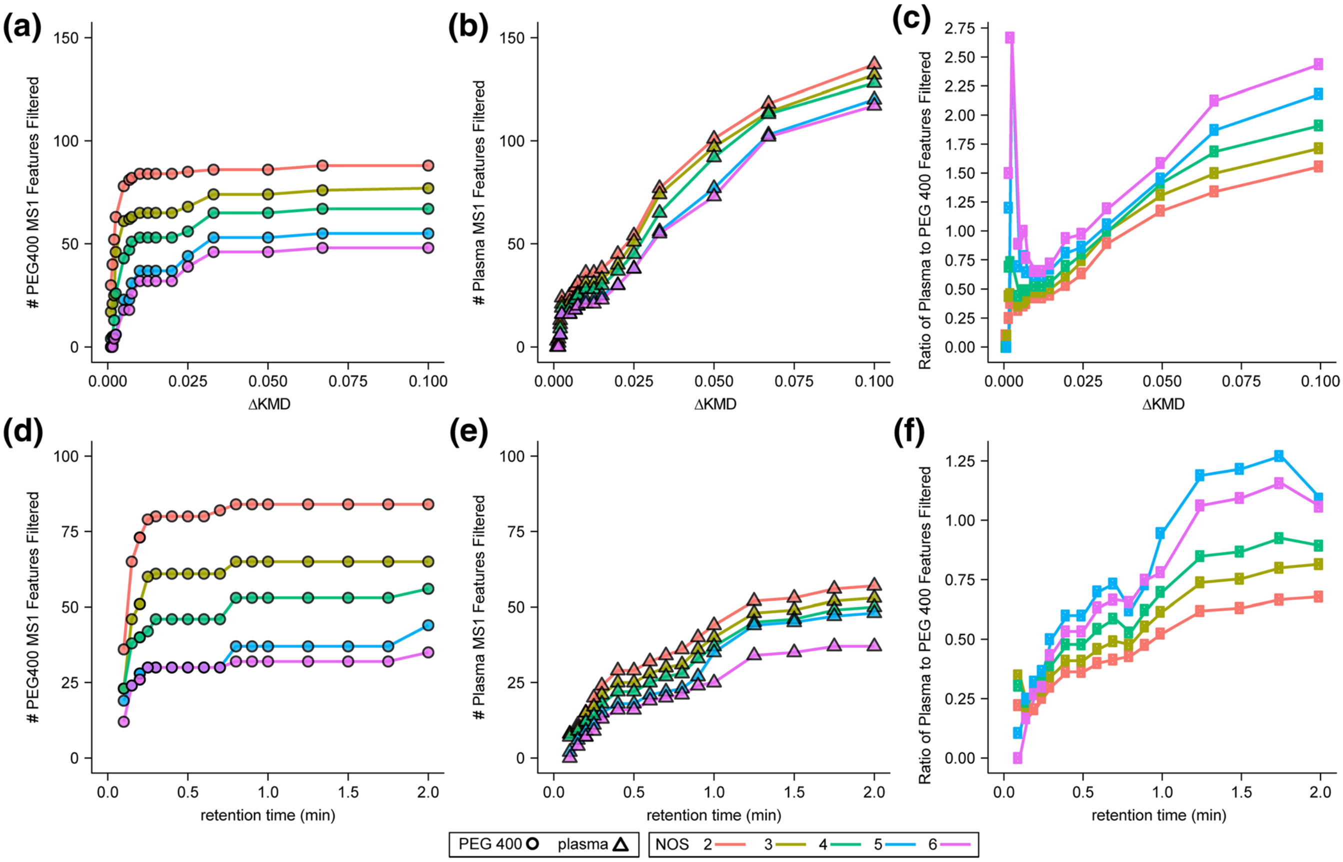


Figure 2. (**a**) ΔKMD versus the number of PEG 400 features filtered (circle) and (**b**) plasma features filtered (triangle). (**c**) ΔKMD versus the ratio of PEG 400 to plasma features filtered. NOS is indicated by color. RT was constant at 0.8 min. (**d**) RT versus the number of PEG 400 features filtered (circle) and (**e**) plasma features filtered (triangle). (**f**) RT versus the ratio of PEG 400 to plasma features filtered. NOS is indicated by color. ΔKMD was constant at 0.01. Data plotted were acquired at a resolution of 17,500

desired signals increases as RT increases. Increasing the NOS parameter reduces the overall number of features filtered in PEG 400 and plasma, Fig. 2f, similar to the behavior observed with the ΔKMD parameter.

The ΔKMD values tested mirror the range of *m/z* accuracy from high to low mass resolution analyzers, e.g., Orbitrap and ToF to linear ion traps to quadruples. The results of the KMF, and any similar mass defect filter, are dependent on *m/z* accuracy. Mass drift, space charge, and peak symmetry could all influence KMD, not evaluated extensively here. However, the effect of acquiring data at different mass resolutions (17,500 and 70,000 using an Orbitrap mass analyzer) was briefly explored. The number of PEG 400 features filtered was similar between data collected at 17,500 and 70,000; however, the number of plasma features filtered was generally less at 70,000 compared to 17,500. This behavior is believed to be attributed to the improved mass accuracy in data acquired at greater mass resolution. The full extent of different data acquisition parameters was not tested in this work.

The data which results from KMF of PEG 400 (Fig. 3a–e) and plasma **(**Fig. 3f–j) collected at a resolution of 17,500 are displayed; KMF results for data collected at a resolution of 70,000 can be found in Fig. S4. The Kendrick mass plot for PEG 400, Fig. 3a, displays the MS^1 features originally present (black) and those which were filtered (red). A large number of the horizontally aligned features within a narrow Kendrick mass defect, suggesting homologous molecules, were filtered. The majority of the MS^1 features which we identified as PEG 400 oligomers (MS^2 supporting putative identification is shown in Fig. S5), eluting between 2 and 4 min could be removed using the defined parameters, Fig. 3b. The MS^1 feature spectrum, i.e., plot of all *m/z* values and their corresponding abundance regardless of their retention time, prior to application of the KMF is displayed in Fig. 3c. The spectrum of MS^1 features retained (those not filtered), Fig. 3d, indicated near complete removal of PEG 400 oligomer signals. The spectrum of MS^1 features filtered by the KMF is displayed in Fig. 3e which contains many oligomer signals present in the original spectrum. Upon closer inspection of the *m/z* differences of signals which were not filtered in the PEG 400 sample, differences between the apparent oligomer peaks did not match the Δ *m/z* requirement to be filtered. The MS^1 feature finding algorithm can cause the *m/z* to fluctuate, by averaging the *m/z* over the aligned samples, and is believed to be the origin of the observed mass difference rather than *m/z* measurement error. The filtering of plasma was minimal, a desired result in this instance, as evident in the Kendrick mass plot (Fig. 3f), MS^1 feature plot (Fig. 3g), and the MS^1 feature spectrum prior to filtering (Fig. 3h), MS^1 features retained after filtering (Fig. 3i), and spectrum MS^1 features removed by the KMF (Fig. 3j).

In addition to visual inspection of the filtered peaks as displayed in Fig. 3 and parameter optimization, we recommend the use of Kendrick mass plots and fractional base units as described in Fouquet and Sato which can improve the visual resolving power of oligomeric series [4]. The KMF using fractional base units is available in the supplementary code (https://github.com/DorresteinLaboratory/Kendrick_Mass_

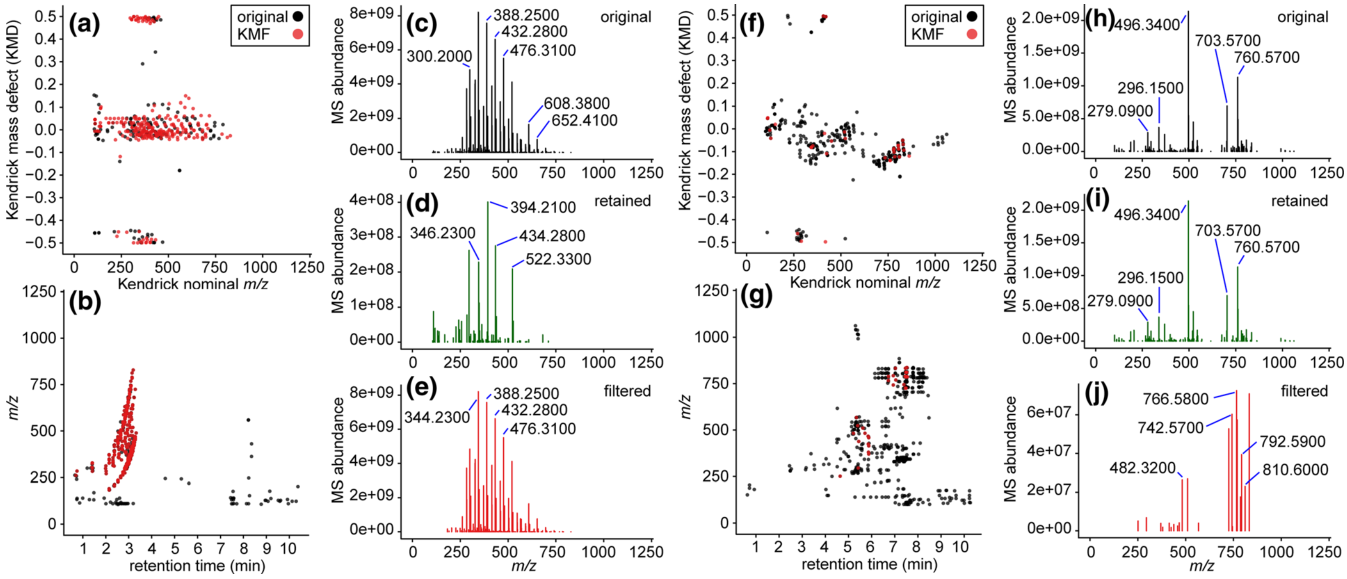


Figure 3. (**a**) Kendrick mass plot (original MS^1 features, black, and filtered MS^1 features, red) of PEG 400. (**b**) MS^1 feature plot of PEG 400, original MS^1 features (black), and MS^1 features filtered (red). (**c**) MS^1 feature spectrum for PEG 400 prior to KMF. (**d**) MS^1 feature spectrum of retained features (not filtered) and (**e**) MS^1 feature spectrum of MS^1 features removed via KMF. (**f**) Kendrick mass plot (original MS^1 features, black, and filtered MS^1 features, red) of plasma. (**g**) MS^1 feature plot of plasma original MS^1 features (black) and MS^1 features filtered (red). (**h**) Illustrative plasma MS^1 feature spectrum prior to KMF; (**i**) plasma MS^1 feature spectrum of MS^1 features retained; and (**j**) MS^1 feature spectrum of MS^1 features removed via KMF. KMF parameters: ΔKMD = 0.01, RT = 0.8 min, and NOS = 2. PEG was filtered using the ethylene oxide unit repeat (*m/z* 44.0262). Data shown were acquired at a resolution of 17,500

Filter) allowing one to quickly change the fractional base unit and create plots. Another visual inspection tool available in the supplementary code is the Gaussian shape tool that takes advantage of the Gaussian shape of most oligomeric series.

Testing of KMF for the Removal of PEG in NIST Plasma Standard Reference Material Spiked with a Swab Extract

An aliquot of plasma was spiked with a swab extract, which was known to contain polymers, and measured in triplicate. This sample illustrates a scenario in which complex undesired chemical background is present. The MS^1 spectra indicated polymer ions, primarily PEG oligomers, which was confirmed by MS^2, Fig. S6. The KMF parameters chosen previously, maximizing the filtering of PEG 400 while minimizing the number of plasma features filtered, were used (ΔKMD = 0.01, RT = 0.8 min, and NOS = 2). The Kendrick mass plot, Fig. 4a, and MS^1 feature plot, Fig. 4b, display a large number of features, which met the KMF criteria. The original MS^1 feature spectrum, Fig. 4c, is composed predominantly of signals, which appear to originate from plasma. However, a large number of peaks at approximately 10% of the base peak appeared to be oligomers (based on equal spacing). Those apparent polymer features are reduced in the KMF filtered MS^1 feature spectrum (Fig. 4d). The KMF features which were removed, plotted in Fig. 4e, indicated a large number of low abundance ions. Incidentally, one abundant ion at *m/z* 496.3400 which is not believed to be an oligomer signal was filtered.

KMF of Axilla Skin Swab Samples in Organ Transplant Cohort Reduces Spectral Complexity Associated with Uncontrolled Deodorant Use

Untargeted metabolomics analysis performed on human skin samples from organ transplant patients on immunosuppressive therapy (*n* = 302), sampled using moistened cotton swabs, were processed using the KMF. One initial question posed was whether the endogenous metabolomic information acquired from hand, face, and axillary skin samples were different. PCA was performed on the untargeted metabolomic data after row sum normalization and pareto scaling. The PCA score plot, principal component 1 (i.e., PC1) vs principal component 2, for the original data is plotted in Fig. 5. The molecular differences detected in hand, face, and axilla samples resulted in only very moderate separation of hand, face, and axillary samples (red, green, and black, respectively). The dispersion of sample points

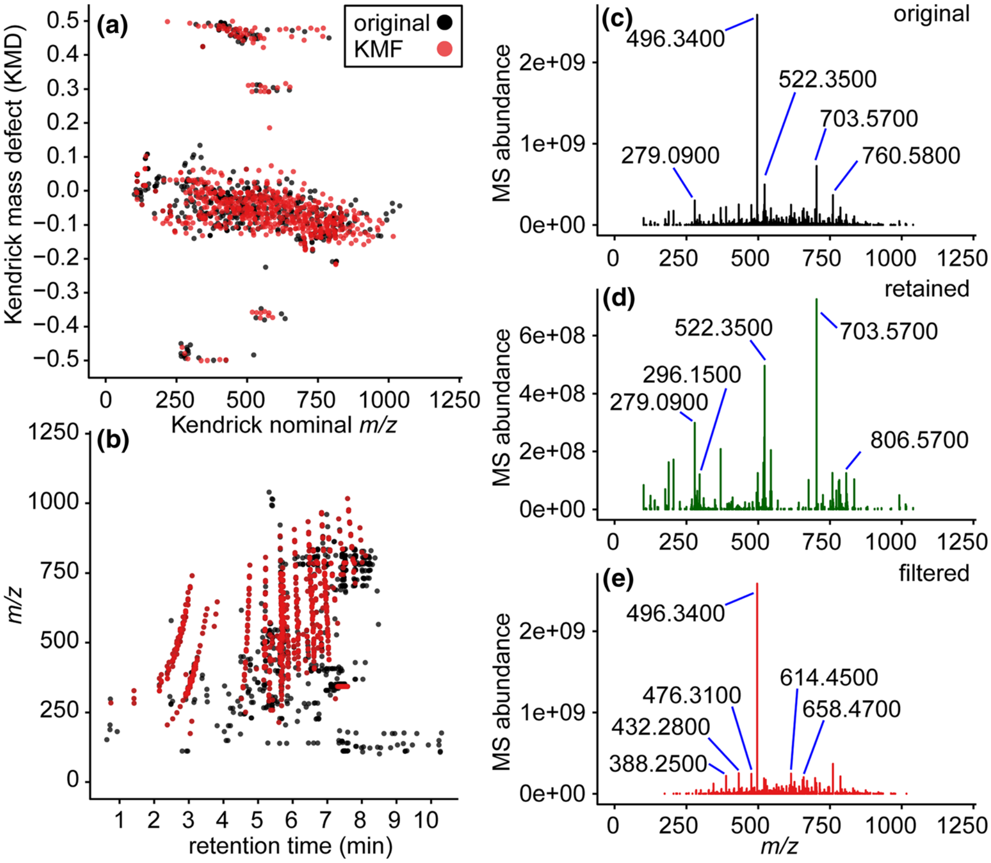


Figure 4. (**a**) Kendrick mass plot (original MS^1 features, black, and filtered MS^1 features, red) of plasma spiked with swab extract filtering. (**b**) MS^1 feature plot of plasma spiked with swab extract, original MS^1 features (black) and MS^1 features filtered (red). (**c**) Plasma spiked with swab extract prior to KMF, (**d**) MS^1 feature spectrum of features retained, and (**e**) MS^1 features removed via KMF. KMF parameters: ΔKMD = 0.01, RT = 0.8 min, and NOS = 2. PEG was filtered using the ethylene oxide unit repeat (*m/z* 44.0262). Data shown were acquired at a resolution of 17,500

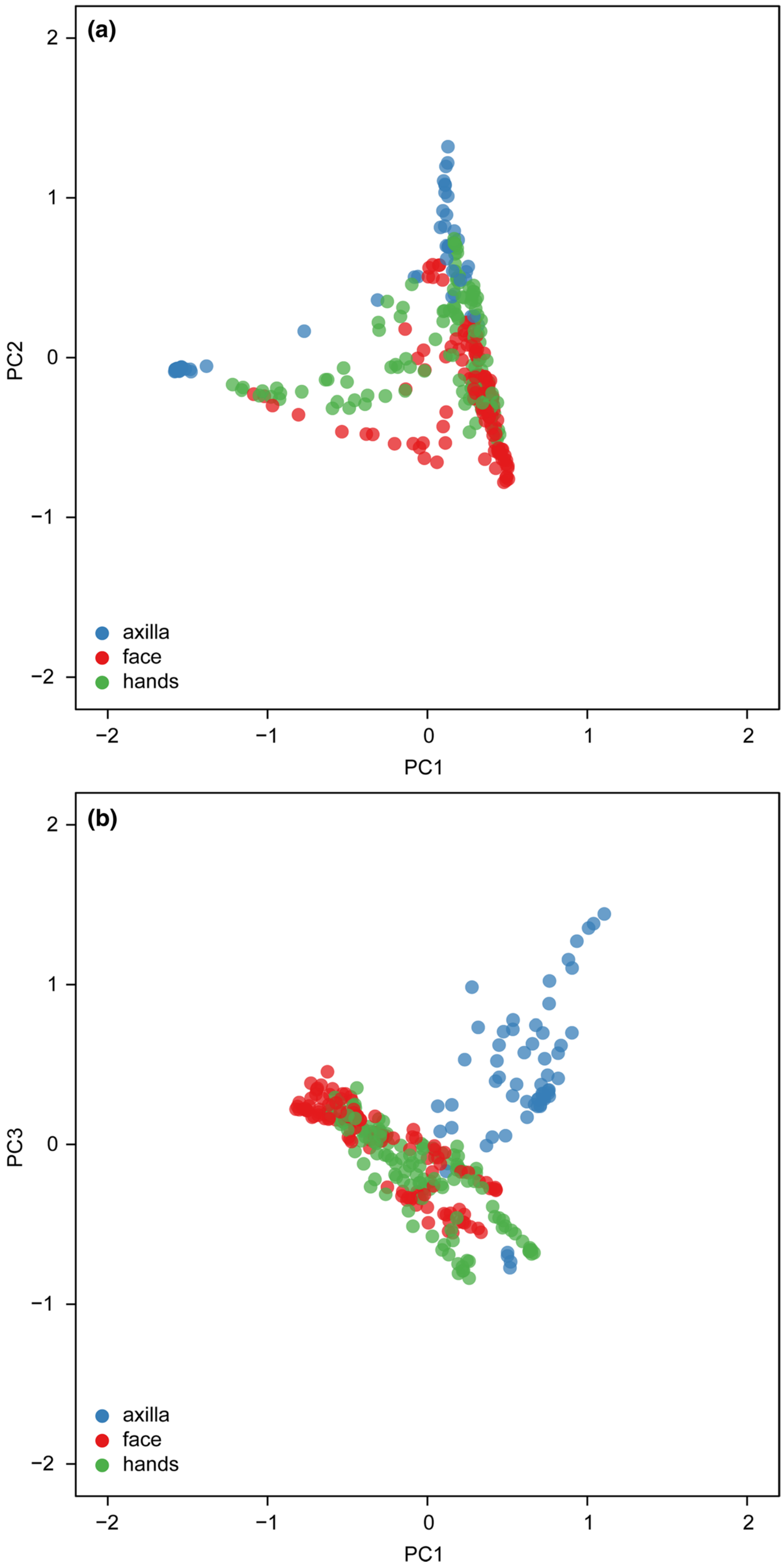


Figure 5. (**a**) PCA score plot of original data obtained from the skin swab samples, hands (green), face (red), and axilla (blue), of organ transplant recipients, 15 subjects. (**b**) PCA score plot of KMF data obtained from the skin swab samples, hands (green), face (red), and axilla (blue), of organ transplant recipients, 15 subjects. KMF was performed for PPG. KMF parameters: ΔKMD = 0.01,RT = 0.8 min, and NOS = 2

in the score plot, Fig. 5a, along PC1 (the PC of greatest contribution to data variance; 8.6%) was investigated and found to be partly due to the presence of polypropylene glycol (PPG) oligomer ions. Additional PCA score plots and loading values can be found in the supplementary information (Fig. S7, Table S1, and electronic supplementary information—Table S2). The PPG oligomer signals were characteristically separated by *m/z* 58.0419 (an illustrative example is shown in Fig. 6c). These peaks were detected in the axilla samples from a number of, but not all, subjects. We hypothesize that the presence of PPG is related to deodorant use as PPG is present in the formulation of deodorant and other skin care products.

In this instance, the removal of the variance due to the PPG ions in the data was desired in order to better visualize the compositional differences between hand, face, and axillary samples independent of deodorant use. The KMF was adjusted for PPG, i.e., Kendrick mass scaling and the specification of the integer unit repeat. The parameters previously used for the filtering of PEG were applied here, i.e., ΔKMD of 0.01, RT of 0.8, and NOS of 2. The Kendrick mass plot and MS^1 feature plot are displayed in Fig. 6a, b. respectively. A large number of features meet the KMF filtering criteria, which while not surprising in this case illustrates how many MS^1 features can be linked to polymers and how such information can complicate interpretation. The KMF removed singly- and doubly-charged species (Fig. S8). The *m/z* difference and homology are preserved between the n^{th} oligomer peak equal to the charge in the oligomer distribution, for example the 2nd consecutive PPG oligomer from any signal in a doubly-charged distribution would meet the Δ*m/z* 58 filtering requirement for PPG. The accuracy of filtering for multiple charged ions was not evaluated in this work. A representative MS^1 feature spectrum from subject BF1637, right axillary, is displayed prior to KMF in Fig. 6c. The MS^1 feature spectrum of retained features is shown in Fig. 6d. The MS^1 feature spectrum displaying the features filtered is plotted in Fig. 6e which clearly indicates that the majority of PPG features in this PPG dominated example were removed.

PCA was performed on the KMF data, Fig. 5b. The differential grouping of axilla samples compared to skin (hands and face) is more apparent, and the PCA loadings indicate that the largest source of variance, PC1 (5.0%), no longer is associated with PPG oligomers peaks (PCA loadings are tabulated in Table S2 and electronic supplementary information—Table S4). PCA score plots visualizing combinations of different PCs can be found in

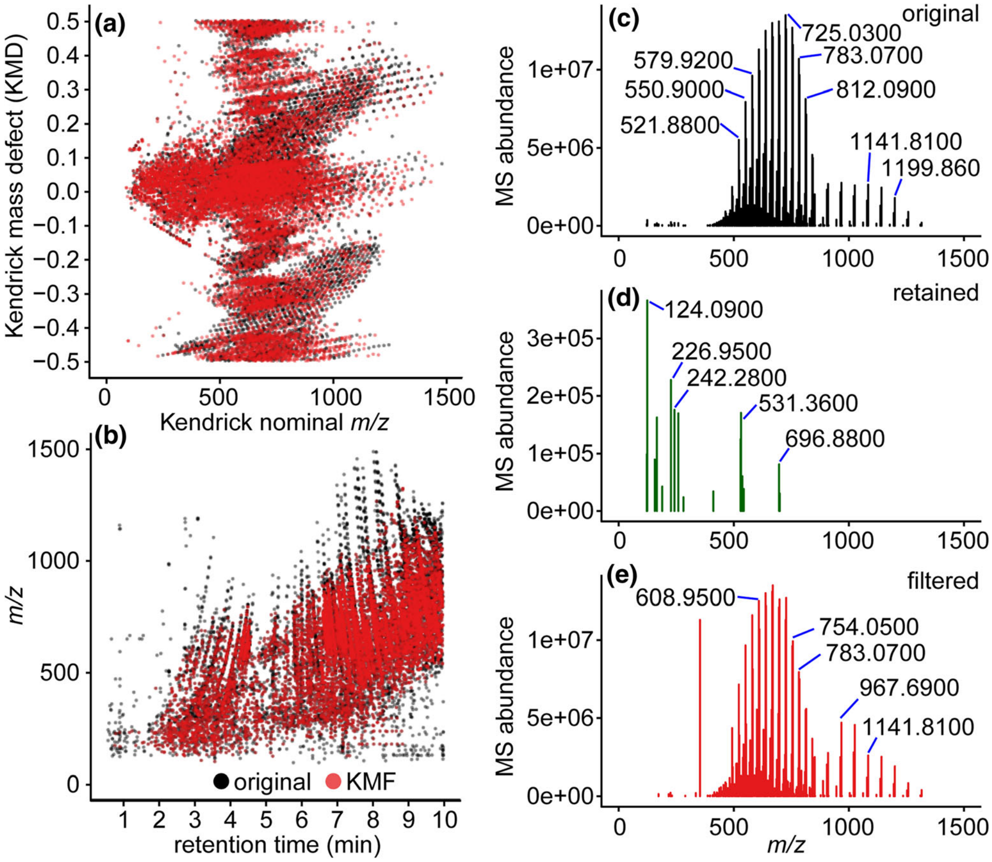


Figure 6. (**a**) Kendrick mass plot (original MS^1 features, black, and filtered MS^1 features, red) of skin swab samples of organ transplant recipients for PPG (original feature, black, and filtered features, red). (**b**) MS^1 feature plot, original MS^1 features (black) and MS^1 features filtered (red). (**c**) Illustrative MS^1 feature spectrum of an axilla sample from subject BF1637, right axillary, believed to contain PPG due to deodorant use. (**d**) MS^1 feature spectrum of features retained after KMF and (**e**) MS^1 feature spectrum of MS^1 features removed via the KMF. KMF parameters: ΔKMD = 0.01, RT = 0.8 min, and NOS = 2. PPG was filtered using the propylene oxide unit repeat (*m/z* 58.0419)

the Supplementary Information (Fig. S9). Note, differences in the PCA score plot are anticipated when changing the number of variables. This real-world experiment exemplifies a situation in which control over the polymer content would have been problematic without control of deodorant use a priori, and one in which a large number of different polymer sources are possible, which makes proper controls difficult to obtain. The KMF processing of data improved interpretation and investigation of the initial question of differentiating hands, face, and axillary samples based on the endogenous metabolome. However, it should be noted that filtering of the data to remove polymer ions does not compensate for the signal suppression caused by polymers during electrospray ionization which will influence the ions observed as well as abundance measurement.

KMF for Exploring Data Compositionality

A priori knowledge of undesired chemical background is rare, and the KMF is particularly suitable in such scenarios (e.g., unknown or unanticipated source of background). In addressing the lack of a priori knowledge, we applied the KMF to explore data compositionality prior to the filtering step. Compositional analysis in this manner is intended to be only informative; the performance of evaluating the accuracy of selection was not evaluated in this study. As the combinatorial nature of the filtering criteria can generate false positives, manual inspection of selected signals is recommended.

The data compositionality of plasma sample spiked with a swab extract ($n = 3$), previously discussed when optimizing KMF parameters, is displayed in Fig. 7. The evaluated background ions were split into the following categories: composition, containment, and source. Note, the evaluated background was not comprehensive, but users can add background signals of interest to the freely available code. The contaminant category included common background polymers and signals including perfluorinated molecules (unit repeat of CF_2), polysiloxanes, PPG (unit repeat of $C_3H_6O_1$), and PEG (unit repeat of $C_2H_4O_1$). Compositional analysis via KMF of the plasma sample spiked with a swab extract, Fig. 7, indicated the presence of a large number of PEG MS^1 features matching expectation. Few features meet the criteria for other polymers. The composition category included unit repeats with masses associated with CH_2, C_2H_4, C_3H_6, C_4H_8, and O, similar to traditional uses of Kendrick mass analysis of data. MS^1 features associated with the composition category are only informative in this instance and should not be filtered, but reveal high-level information on sample components. Similarly, the source category is largely informative in this instance, but could be valuable in understanding ionization generated signals, intentionally or inadvertently. The accuracy in determining the presence of MS^1 features belonging to the source category is beyond the scope of this publication, but the aim of future studies.

Conclusion

Mass spectrometry, particularly untargeted analysis, frequently encounters some degree of chemical background such as polymers. Systematic investigation of the KMF user-defined parameters, ΔKMD, RT, and NOS, was performed using a PEG 400 standard and plasma. Testing of the KMF parameters chosen to maximize PEG 400 and minimize plasma feature removal were applied to plasma spiked with a swab extract, illustrative of a situation in which complex undesired chemical background is present. The MS^1 features from the plasma spiked with a swab extract which were not filtered in the MS^1 feature plot were highly reminiscent of those in the plasma standard MS^1 feature plot. The skin samples obtained from the organ transplant recipients illustrate the effects of uncontrolled polymer background on multivariate statistical analysis—the observation of which is likely amplified by the low biomass skin samples analyzed—and how the KMF can be used to remove such interferences and clarify interpretation. The KMF was also used to explore the composition of the data acquired providing a

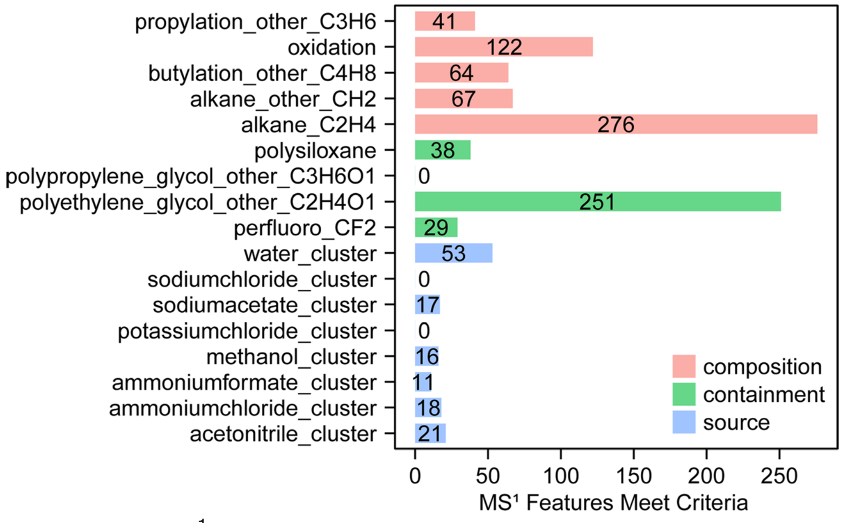


Figure 7. Compositional analysis. The MS^1 features which meet KMF criteria associated with the composition, (red) containment, (green) and source (blue) categories are plotted

high-level overview of MS^1 features, which display unit repeat homology without a priori knowledge. Such an analysis might serve as a first step to determine which signals are most abundant and the target of subsequent filtering.

The parameters in this instance were determined by maximizing the filtering of PEG 400 MS^1 features while minimizing NIST plasma SRM. In general, parameters should be selected based on the data and weighing desired outcomes (e.g., the extent of filtering against inappropriate removal of non-polymer ions). Ideally, the best use case requires a pilot experiment in which the desired polymer can be spiked into the matrix allowing the ideal parameters to be determined; this process is usually not feasible due to time and resource constraints. Given the constraints, we recommend that users run the filter on a few representative samples, and use the plots provided in the supplementary Jupyter notebook to adjust the parameters and inspect the filtering results. Confirmation of MS^1 features which are retained or filtered is recommended. We further recommend the use of GNPS [13] for MS^2 queries for putative metabolite identification and contribution of background MS^2 spectra into the public spectra library.

The KMF is not intended to replace appropriate study design and careful sample handling. It is intended to be used in situations in which other options, e.g., background subtraction, are not possible or ineffective. The proposed KMF can theoretically be applied to mass spectral data obtained using any mass analyzer; however, mass accuracy will influence filtering. Additionally, the method by which MS^1 feature finding is performed will influence filtering quality. We anticipate that the KMF could also be used when chemical separation is not performed, e.g., matrix-assisted laser desorption, nanoelectrospray, and ambient ionization techniques. We envision the use of the KMF to select only polymer ions in MS^1 data, filtering all non-polymer MS^1 features, in order to study the polymer content of samples.

Acknowledgements

The authors would like to acknowledge the scientific support of the Collaborative Mass Spectrometry Innovation Center and the Center for Microbiome Innovation, University of California, San Diego. Funding support for this work is provided by the US National Institutes of Health 1R03CA211211-01 and 5P41GM103484-07.

References

1. Kendrick, E.: A mass scale based on CH2 = 14.0000 for high resolution mass spectrometry of organic compounds. Anal. Chem. **35**, 2146–2154 (1963)
2. Hughey, C.A., Hendrickson, C.L., Rodgers, R.P., Marshall, A.G., Qian, K.: Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal. Chem. **73**, 4676–4681 (2001)
3. Kim, S., Kramer, R.W., Hatcher, P.G.: Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the Van Krevelen diagram. Anal. Chem. **75**, 5336–5344 (2003)
4. Fouquet, T., Sato, H.: Extension of the Kendrick mass defect analysis of Homopolymers to low resolution and high mass range mass spectra using Fractional Base units. Anal. Chem. **89**, 2682–2686 (2017)
5. Sleno, L.: The use of mass defect in modern mass spectrometry. J. Mass Spectrom. **47**, 226–236 (2012)
6. Zhang, H., Zhang, D., Ray, K., Zhu, M.: Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. J. Mass Spectrom. **44**, 999–1016 (2009)
7. McMillan, A., Renaud, J.B., Gloor, G.B., Reid, G., Sumarah, M.W.: Post-acquisition filtering of salt cluster artefacts for LC-MS based human metabolomic studies. J. Cheminform. **8**, 44 (2016)
8. Paguigan, N.D., El-Elimat, T., Kao, D., Raja, H.A., Pearce, C.J., Oberlies, N.H.: Enhanced dereplication of fungal cultures via use of mass defect filtering. J. Antibiot. (Tokyo). **70**, 553–561 (2017)
9. Simón-Manso, Y., Lowenthal, M.S., Kilpatrick, L.E., Sampson, M.L., Telu, K.H., Rudnick, P.A., Mallard, W.G., Bearden, D.W., Schock, T.B., Tchekhovskoi, D.V., Blonder, N., Yan, X., Liang, Y., Zheng, Y., Wallace, W.E., Neta, P., Phinney, K.W., Remaley, A.T., Stein, S.E.: Metabolite profiling of a NIST standard reference material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. Anal. Chem. **85**, 11725–11731 (2013)
10. Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P.: A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. **30**, 918–920 (2012)
11. Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M.: MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. **11**, 395 (2010)
12. Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J.: pcaMethods a bioconductor package providing PCA methods for incomplete data. Bioinformatics. **23**, 1164–1167 (2007)
13. Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W.-T., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., P, C.A.B., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., Bandeira, N.: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. **34**, 828–837 (2016)