metabolites

MDPI

Article

Challenges in Lipidomics Biomarker Identification: Avoiding the Pitfalls and Improving Reproducibility

Johanna von Gerichten [1,†], Kyle Saunders [1,†], Melanie J. Bailey [1], Lee A. Gethings [2], Anthony Onoja [3], Nophar Geifman [3] and Matt Spick [3,*]

[1] School of Chemistry and Chemical Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, UK; j.vongerichten@surrey.ac.uk (J.v.G.); ks00916@surrey.ac.uk (K.S.); m.bailey@surrey.ac.uk (M.J.B.)
[2] Waters Corporation, Wilmslow SK9 4AX, UK; lee_gethings@waters.com
[3] School of Health Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford GU2 7XH, UK; a.onoja@surrey.ac.uk (A.O.); n.geifman@surrey.ac.uk (N.G.)
* Correspondence: matt.spick@surrey.ac.uk
† These authors contributed equally to this work.

Abstract: Identification of features with high levels of confidence in liquid chromatography–mass spectrometry (LC–MS) lipidomics research is an essential part of biomarker discovery, but existing software platforms can give inconsistent results, even from identical spectral data. This poses a clear challenge for reproducibility in biomarker identification. In this work, we illustrate the reproducibility gap for two open-access lipidomics platforms, MS DIAL and Lipostar, finding just 14.0% identification agreement when analyzing identical LC–MS spectra using default settings. Whilst the software platforms performed more consistently using fragmentation data, agreement was still only 36.1% for MS^2 spectra. This highlights the critical importance of validation across positive and negative LC–MS modes, as well as the manual curation of spectra and lipidomics software outputs, in order to reduce identification errors caused by closely related lipids and co-elution issues. This curation process can be supplemented by data-driven outlier detection in assessing spectral outputs, which is demonstrated here using a novel machine learning approach based on support vector machine regression combined with leave-one-out cross-validation. These steps are essential to reduce the frequency of false positive identifications and close the reproducibility gap, including between software platforms, which, for downstream users such as bioinformaticians and clinicians, can be an underappreciated source of biomarker identification errors.

Keywords: lipidomics; separation science; mass spectrometry; bioinformatics; machine learning; retention time

Citation: von Gerichten, J.; Saunders, K.; Bailey, M.J.; Gethings, L.A.; Onoja, A.; Geifman, N.; Spick, M. Challenges in Lipidomics Biomarker Identification: Avoiding the Pitfalls and Improving Reproducibility. *Metabolites* **2024**, *14*, 461. https://doi.org/10.3390/metabo14080461

Academic Editors: Konstantinos Kouremenos and David J. Beale

Received: 15 June 2024
Revised: 14 August 2024
Accepted: 15 August 2024
Published: 19 August 2024

1. Introduction

The identification of lipids allows for biological interpretation, as well as the association of specific lipids with cellular processes, signaling pathways, and disease conditions [1,2]. In addition, bioinformatics allows for the integration of lipidomic identifications with other omics datasets, such as genomics, proteomics, and metabolomics, to provide a more comprehensive understanding of cellular processes and interactions [3,4]. Therefore, accurate and reproducible identification is critical when searching for biomarkers, features that can indicate the presence or prognostics of a disease, allowing for early diagnosis and personalized medicine. Conversely, inaccurate identification can lead to incorrect conclusions and potentially misleading research findings [5–7].

The desire for accurate identification—as in other areas of omics research—has driven the creation of the Lipidomics Standards Initiative (LSI) [8]. This initiative sets out recommended procedures for quality controls, reporting checklists, and minimum reported

information [9,10]. However, the LSI is less mature in its recommendations and implementation than, for example, the Metabolomics Standards Initiative (MSI), which itself is still evolving [11,12]. The difficulty in defining parameters for confident annotation partly reflects the sheer range of potential lipids and matrices, encompassing cells, biofluids, tissues, plant extracts, and others [13,14]. This range of samples is then multiplied by the panoply of analytical platforms and separation techniques, such as reversed-phase LC or hydrophobic interaction LC (HILIC). This makes standards-based matching and other best practices more challenging, costly, and time-consuming, especially at the discovery stage.

These issues are well described in the literature, but an underappreciated source of reproducibility problems in untargeted analysis is the lack of consistency in outputs from lipidomics software platforms. Whilst analytical chemists specializing in lipids will often be aware of the issues with peak annotation and feature identification, this will not always be the case for users such as bioinformaticians and clinicians. These software solutions typically pre-process lipid spectra in a five-step workflow previously summarized by Song et al. [15], comprising (i) baseline and noise reduction, (ii) peak identification and extraction, (iii) smoothing, (iv) calculation of signal-to-noise ratios, and (v) isotope identification and deconvolution. Following these steps, accurate mass-to-charge ratio (m/z) matching is used for identification, combined with fragmentation spectra derived from MS^2. However, such MS^2 spectra are not infallible given the potential for co-elution of lipids within the precursor ion selection window and co-fragmentation. Furthermore, MS^2 may not be practical for lower-abundance lipids. The same abundance issues hinder ion-mobility mass spectrometry [16], a powerful technique that allows for the separation of isobaric lipids but requires specialist instruments and has trade-offs between sensitivity and resolving power. Inconsistencies can also be driven by the use of different lipid libraries such as LipidBlast, LipidMAPS, ALEX123, and METLIN [17]. These issues can also be magnified by different spectral alignment methodologies, which are often opaque to the end user and can cause substantial differences in peak identification. One inter-laboratory comparison of lipidomics LC–MS alignment found an agreement for post-processed features of around 40% between the two laboratories [18].

Another reason for inconsistencies between platforms in untargeted analyses and the need for outlier detection and manual curation is that the majority of lipidomics software tools do not make full use of retention time (t_R), a rich source of information that has been used extensively in machine learning approaches to improve proteomic identifications [19–22]. Machine learning methods that use algorithms trained on specific columns and operating conditions do not generalize well and are not straightforward to implement across the full range of lipidomics modalities. Whilst the producers of lipidomics software are, of course, aware of these limitations and recommend that putative lipid identifications be manually curated, this can be time-consuming, as well as imposing barriers to entry around lipidomics research for bioinformaticians and clinicians. This is especially the case in secondary analyses such as obtaining validated biomarkers, meta-analyses, or systematic reviews [23,24].

Whilst these issues are recognized by the LSI and other best practice guidelines, there is a paucity of benchmarking of software applications relative to LC–MS methodologies. In this work, we provide a case study in order to reveal the problems researchers face in the form of potentially inaccurate biomarker identifications by processing an identical set of LC–MS spectra using two popular lipidomics platforms: MS DIAL and Lipostar. Whilst workflows for the individual platforms are well-described, to our knowledge, this is the first cross-platform comparison on a single LC–MS dataset. We use this case study to highlight the importance of manual curation and the pitfalls of relying too heavily on 'top hit' software identifications, including for discovery work or where groups do not have in-house lipid curation libraries. This can be particularly relevant to new researchers in the field, who may face challenges in receiving sufficient support for analytical training and education [25]. We also demonstrate a novel data-driven quality control step for outlier detection applicable to any untargeted lipidomics analysis by using support vector

machine regression combined with leave-one-out cross-validation to identify potentially false positive identifications. These types of quality control steps can be performed on computers typically available in a laboratory setting, without recourse to high-performance computing clusters, and can support manual inspection processes. Such additional steps, especially manual curation, are necessary even where MS^2 spectra are used, particularly in instances of conflicting identifications by differing platforms.

2. Materials and Methods

2.1. PANC-1 Lipid Extraction LC–MS Dataset

The lipidomics case study dataset used in this work analyzed a lipid extraction of a human pancreatic adenocarcinoma cell line (PANC-1, Merck, Gillingham, UK, cat no. 87092802). Lipids were extracted by a modified Folch extraction using a chilled solution of methanol/chloroform (1:2 v/v) according to the protocol described by Zhang et al. supplemented with 0.01% butylated hydroxytoluene (BHT) to prevent lipid oxidation [26]. An Avanti EquiSPLASH® LIPIDOMIX® quantitative mass spectrometry internal standard, a mixture of deuterated lipids, was added to the extract, and the resulting mixture was then diluted to 280 cells/µL. The final EquiSPLASH concentration was 16 ng/mL. Injections of 5 µL of the lipid extract were analyzed using an Acquity M-Class UPLC system (Waters, Wilmslow, UK) coupled to a ZenoToF 7600 mass spectrometer (Sciex, Macclesfield, UK) operated in positive mode. A Luna Omega 3 µm polar C18 column was used (50 × 0.3 mm, 100 Å, Phenomenex, Macclesfield, UK, cat no. 00B-4760-AC) for microflow separation at 8 µL/min. A binary gradient was carried out using eluent A (60:40 acetonitrile/water) and B (85:10:5 isopropanol/water/acetonitrile), both supplemented with 10 mM ammonium formate and 0.1% formic acid. Separation was achieved using the following gradient: 0–0.5 min, 40% B; 0.5–5 min, 99% B; 5–10 min, 99% B; 10–12.5 min, 40% B; 12.5–15 min, 40% B. Mass spectrometry settings are set out in Supplementary Material (Table S1). The analysis was conducted in 2023; the untargeted approach described here, including the use of positive mode, was adapted from a commonly used method [27].

The output files were then processed in two lipidomics applications, MS DIAL (v4.9.221218) and Lipostar (v2.1.4) [28,29], using settings set out in full in Supplementary Material (Tables S2 and S3); settings were chosen to make the assumption sets used by the two platforms as similar as possible, but the default libraries were used. For data-driven outlier analysis, a .csv file was prepared for each output, containing the chemical formula for the parent molecule, the class of lipid, the lipid t_R, MS^1 and MS^2 status, and the putative identification. Lipids with t_R below 1 min were considered to have no column retention at all (i.e., eluting with the solvent front) and were excluded from the outlier analysis as having no useful dependent variable.

2.2. Comparison of Outputs

Both Lipostar and MS DIAL produced a list of putative identifications based on both MS^1 and MS^2 data. All lipidomics (and omics software in general) rely on user settings but also built-in analytical steps for alignment and lipid library access, which may produce inconsistencies. The two output datasets from identical input spectral files were compared to identify the overlapping and unique lipid annotations. Lipid identifications were only considered to be in agreement if the formula was identical, the lipid class was identical, and the aligned retention time was consistent within 5 s between MS DIAL and Lipostar.

2.3. Post-Software Quality Control Checks of Data

Post-software quality control steps were then conducted on the assumption that the initial output from the lipidomics software would not represent a 'definitive' ground truth. This step—aiming to provide a method and platform-neutral means of improving confidence in lipid annotations—employed a support vector machine (SVM) regression algorithm using leave-one-out cross-validation (LOOCV) in order to predict lipid t_R [30,31]. The independent variable inputs for the algorithm were the atom count of the parent

lipid (i.e., numbers of carbon, hydrogen, nitrogen, oxygen, or other atoms) and lipid class, including inter alia diglycerides (DGs), triglycerides (TGs), phosphatidylcholines (PCs), and ceramides. t_R was the dependent variable. SVM was chosen for its stability of outputs, ability to deal with multicollinearity, e.g., between carbon and hydrogen atom count (C and H count hereafter), and efficient execution time relative to tree-based algorithms. In addition, given an a priori assumption that the latent variables were linear, using a linear kernel can be preferable to step-function models, which can be more prone to overfitting [32]. Numeric variables were auto-scaled, and categorical variables (lipid classes) were one-hot encoded prior to their inclusion [33]. A linear kernel was used with C = 10, and feature importance was assessed by measuring the explanatory contribution of each variable by SHAP values (SHapley Additive exPlanations), which quantify how much each feature contributes to a model's prediction of the dependent variable, in this case, t_R [34,35]. In some cases, SHAP values for a feature can be misleading due to non-confounding redundancy, where a feature explains t_R but also causally drives another feature included in the model, which in turn also explains t_R [36].

The code was developed in Python using the scikit-learn (v1.3), shap (v0.43.0), and chemparse (v0.1.2) libraries [37–39]. The tqdm (vv4.66.1) library was used to include progress bars for the more time-intensive processes [40]. This code is provided in full as a Jupyter Notebook together with the original raw spectral files and the processed outputs from Lipostar and MS DIAL (as described under Data Availability) and requires no additional software other than a Python environment. The code described in this work was run on a standard Windows PC with a 12th Gen Intel Core i7 CPU paired with 32.0 GB of memory, without employing GPU resources, using the Spyder IDE [41].

As a final step in assessing differences in MS^2 spectra, lipid identifications were then reviewed based on confidence criteria reported by the two software platforms and then manually inspected in SCIEX (v3.0.0.3339), with a particular focus on (i) outliers identified by the SVM with LOOCV algorithm described above and (ii) lipids where MS DIAL and Lipostar provided conflicting identifications in spite of MS^2 fragmentation data being available.

3. Results

3.1. Comparison of MS DIAL and Lipostar Outputs

MS DIAL produced 907 putative lipid identifications across 64 lipid classes from the PANC-1 LC–MS spectra, the most common of which being ceramides (232 identifications across six subclasses), ether PCs (75), and PCs (72). Retention times varied from 1.1 min to 12.4 min, and m/z values ranged from 153.1 to 898.8. Lipostar produced 979 putative lipid identifications across 43 lipid classes, the most common of which were PCs (151 identifications), DGs (130), and ceramides (114). Retention times for the Lipostar-identified peaks varied from 1.1 min to 12.3 min, and m/z values ranged from 177.1 to 889.7.

As a simple measure of agreement and disagreement, Figure 1 shows the common (same formula, t_R, and lipid class) and unique identifications for the MS^1-only features and also the features with MS^2 data. In total, the two platforms generated 1653 unique identifications, of which 231 were common to both platforms, or 14.0%. The breakdown of the lipid classes by matched and unmatched status is shown in Figure 2.

3.2. Data-Driven Investigation of Putative Lipid Identifications

In line with best practice, and given the low numbers of common identifications for the PANC-1 dataset between the two platforms, additional investigations were undertaken. First, both platforms provide a variety of 'scores' to help assess confidence in the annotation. For Lipostar, the overall identification score is based on a weighted average of mass score, isotopic pattern score, and fragment score (which itself is a geometric average of the number of fragments score multiplied by ion intensity score) [28]. For MS DIAL, the overall identification score is calculated as a weighted average of MS^2 similarity, MS^1 similarity, retention time similarity, and isotopic similarity [42]. The two 'scores' are calculated in

different ways; therefore, they should not be directly compared, but as shown in Figure 3, there was no clean 'score' threshold to identify matched/overlapping features versus non-matched features. Whilst the software platforms generally reported a higher 'score' for features identified by both platforms, the overlap was far from perfect.

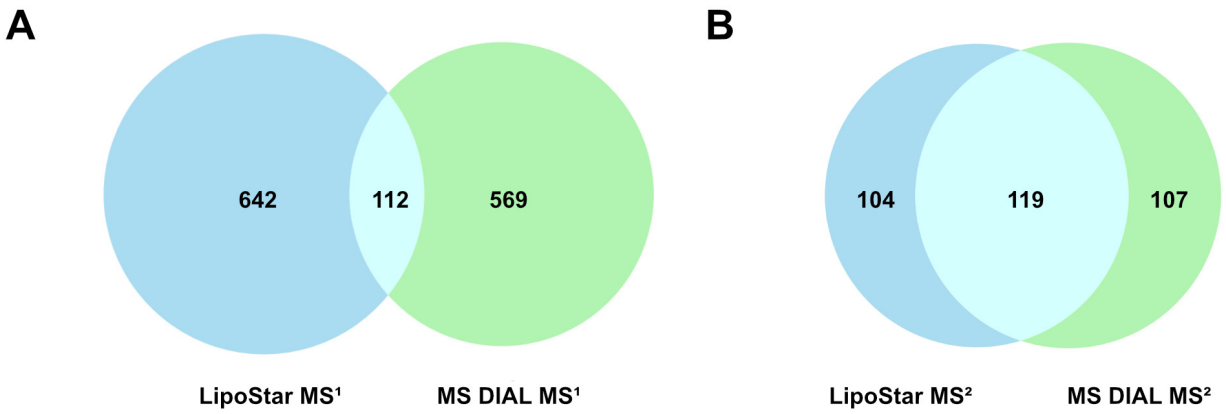

Figure 1. Distinct and overlapping identifications between Lipostar and MS DIAL. (**A**) MS^1 data only and (**B**) MS^2 data only.

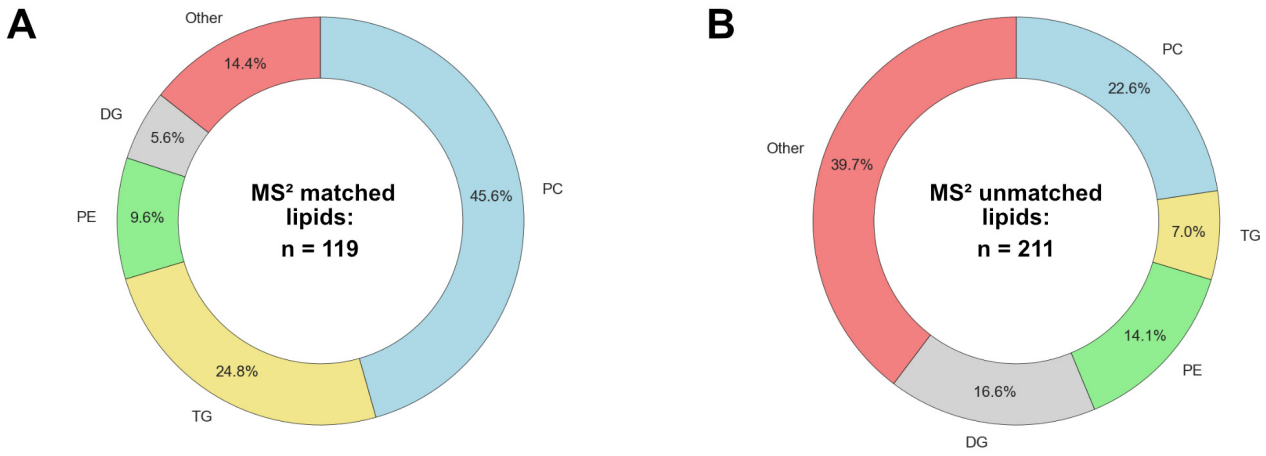

Figure 2. Breakdown of lipid classes identified: (**A**) common MS^2 identifications and (**B**) unique MS^2 identifications.

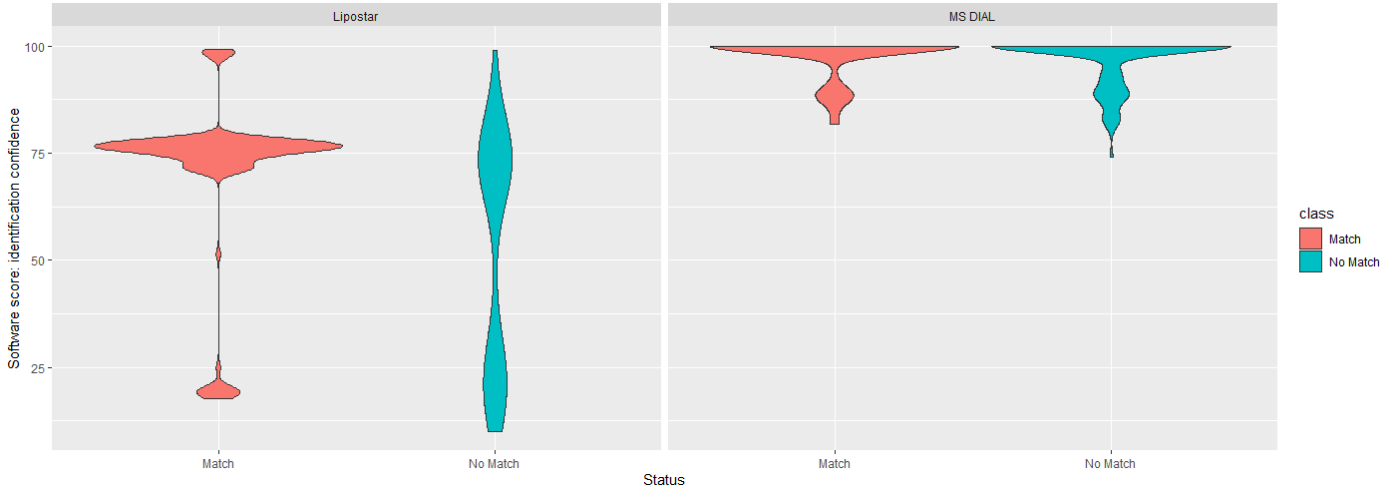

Figure 3. Violin plots for annotation confidence scores of matching and non-matching lipid identifications for Lipostar and MS DIAL. Individual scores range from 0 to 100. MS^2 identifications are only shown here.

Next, a platform-independent data-driven approach was adopted by assessing the overall internal consistency of the elution order of the lipid annotations produced by each software package to identify outliers versus expected t_R values. This was performed without reference to external data, such as custom libraries. The analysis was performed using SVM regression combined with LOOCV, which was applied only to the data internal to the LC–MS spectra analyzed. For the MS DIAL dataset, the algorithm assessed the identifications as being 80.1% internally consistent (i.e., with 19.9% of identifications being further than 5% of the LC–MS runtime from their predicted values). For the Lipostar dataset, the algorithm assessed the identifications as being 69.1% internally consistent, i.e., the lipid identifications showed less internal consistency than with MS DIAL. As a simple metric for overall identification performance, these percentages were consistent with the two platforms being unable to fully validate each others' identifications (Figure 1B).

The Python code used for this data-driven approach also generated a number of visualizations to support post-software investigation of annotations. These are illustrated in Figure 4 for the MS DIAL dataset. The outlier algorithm identified three categories of outlier lipids where t_R was potentially inconsistent with the lipid identification provided by the software. The first category of outliers included cases where lipid identifications were outside the range of the vast majority of lipids in their class, given comparable carbon counts. The process of identifying clear outliers can be seen by a simple visual comparison of actual t_R versus the SVM-predicted t_R in Figure 4A,B.

The second category of outliers was unexpected variations in elution time in sequences of double bonds. An example is shown in Table 1, where decreasing saturation (increasing hydrogen count) of TGs was associated with increasing t_R. However, in the case of $C_{51}H_{98}O_6$, this relationship was inconsistent. The third category of outliers is related to head group ordering. PCs are formed of a quaternary charged amine and two fatty acid chains. DGs are formed of glycerol with two fatty acid chains. The charged quaternary amine renders PCs more hydrophilic than DGs, and so on a C18 column, PCs with equivalent fatty acid chains would be expected to elute earlier than DGs, not at the same time; an example is shown in Table 1.

Table 1. Example lipids flagged for review versus those with an internally consistent t_R—MS DIAL dataset.

MS DIAL Identified Features	Actual t_R (min)	Predicted t_R (min)	Δ (min)
Inconsistencies in t_R: saturation			
TG $C_{51}H_{92}O_6$	7.42	7.49	0.07
TG $C_{51}H_{94}O_6$	7.55	7.57	0.02
TG $C_{51}H_{96}O_6$	7.69	7.65	−0.04
TG $C_{51}H_{98}O_6$	**6.62**	**7.74**	**1.12**
Inconsistencies in t_R: headgroups			
DG $C_{36}H_{61}D_7O_5$	6.55	6.52	−0.03
PC $C_{36}H_{64}NO_8P$	**6.50**	**5.42**	**−1.08**

Bold text indicates >5% run-time Δ, flagged for review.

SHAP values were used to assess feature importance for the outlier detection algorithm and are summarized in Figure 4D. SHAP value beeswarm plots show the contribution of each variable (C count, headgroup, etc., on the *y*-axis) to each individual forecast of t_R for each lipid (the model output on the *x*-axis) and provide more individual detail than a plot of overall feature importances. H count was the most explanatory variable, with higher-than-average numbers of H atoms (red) associated with increased model output (predicted t_R). H count was selected by the model over C count, but it should be noted that this is an example of non-confounding redundancy, as both H and C are causally driven by the same latent variable, in this case, the length of acyl chains. Given the information already available from the H count, in practice, increased C count for a given value of H reduced

predicted t_R slightly. This represents the saturation latent variable (i.e., changes in the CH relationship). The SHAP values also show the impact of the headgroup independently—for example, if all other variables were held equal, a PC headgroup would reduce t_R.

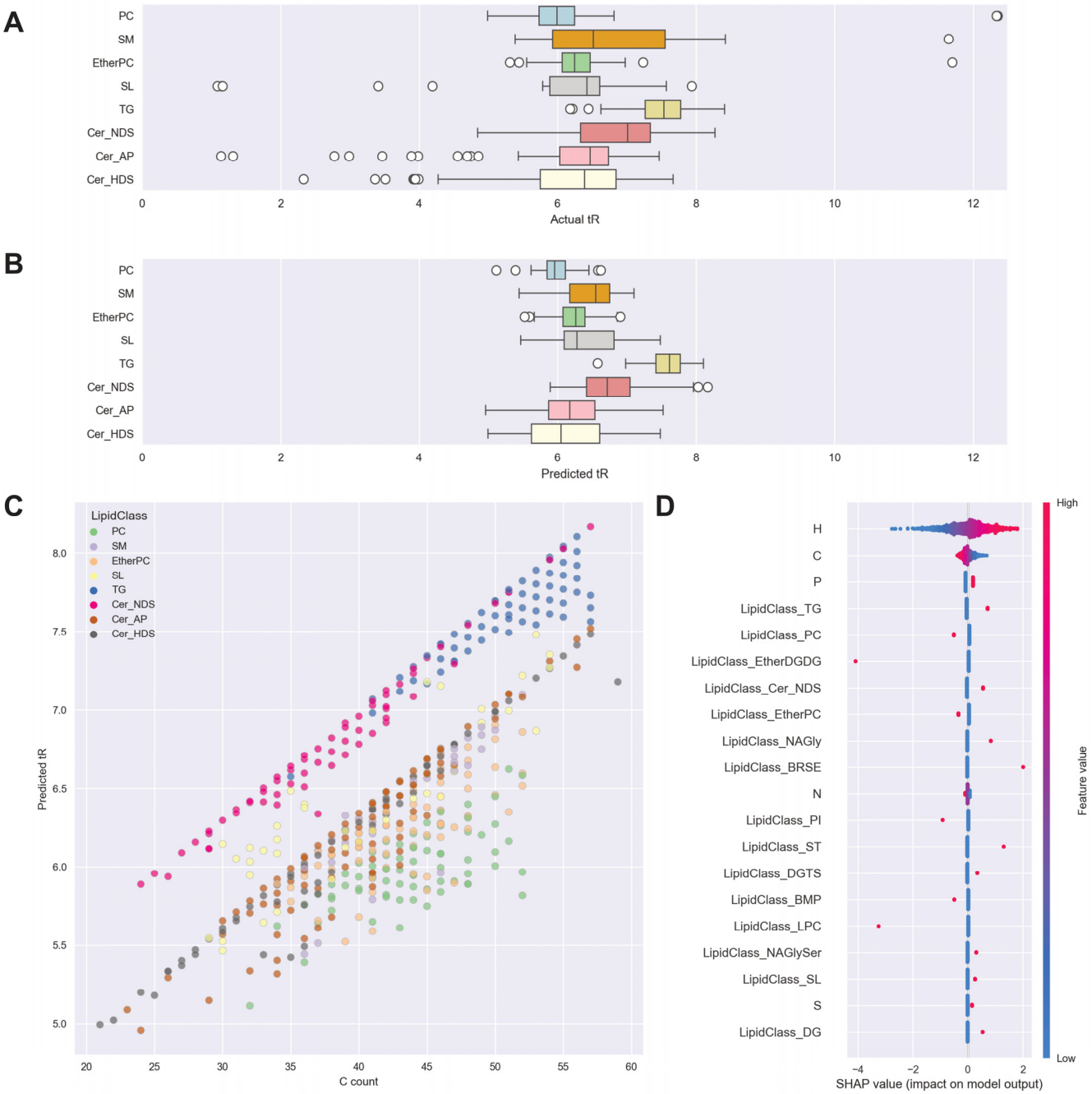

Figure 4. (**A**,**B**) Actual and predicted t_R plotted as boxplots by lipid class: 8 most abundant lipid classes shown. The upper and lower bounds of boxes show the interquartile range. (**C**) predicted t_R plotted against carbon atom count, with lipid class shown by color, for 8 most abundant lipid classes. (**D**) SHAP value beeswarm plots for feature importance, atom count, and 8 most abundant lipid classes shown—*x*-axis represents the impact on model output (predicted t_R), each dot represents a sample. The color scale indicates the impact of feature value, where red is an above-average value for a feature such as carbon count, and blue is a below-average value for a feature. TG—triglycerides, SM—sphingomyelins, SL—sphingolipids, PC—phosphatidylcholines, Cer—ceramides including alpha-hydroxy ceramides (Cer_AP), non-hydroxy ceramides containing dihydrosphingosine (Cer_NDS), and hydroxy ceramides containing dihydrosphingosine (Cer_HDS).

3.3. Manual Investigation of Putative Lipid Identifications

Following the data-driven steps described in the preceding section, MS^2 spectra for the lipids were investigated, paying particular attention to any flagged by the SVM outlier analysis or to those with inconsistent identifications between MS DIAL and Lipostar. A number of lipid identifications were found to have MS^2 spectra inconsistent with the putative software identifications, falling under two headings.

Co-elution problems: where several lipids elute at the same time (i.e., within the precursor ion selection window), a variety of lipid fragments may be present in the MS^2 spectra. An example of this is shown in Table 2 at t_R 6.78 and 6.80. MS DIAL identified one DG, one TG, and two ceramides. At the same t_R, Lipostar identified three DGs, one TG, and one ceramide. Manual inspection of the MS^2 spectra (shown in Supplementary Materials, Figure S1) indicated that the four MS DIAL identifications were correct, and so also were three of the Lipostar identifications, i.e., both platforms missed lipids identified by the other platform. Lipostar additionally generated two identifications that could not be validated by manual inspection.

Table 2. Conflicting identifications: MS DIAL versus Lipostar.

MS DIAL Identified Features	t_R (min)	Lipostar Identified Features	t_R (min)
Identification problems: co-elution of lipids			
DG 34:0 | DG 16:0_18:0	6.78	DG (15:0/16:0/0:0)	6.78
TG 41:1;O | TG 9:0_17:0_15:1;O	6.78	DG (15:1/18:1/0:0)	6.78
Cer 42:2;O2 | Cer 18:1;O2/24:1	6.80	TG (13:0/13:0/16:0)	6.78
Cer 42:2;O2 | Cer 18:1;O2/24:1	6.80	DG (15:0/18:1/0:0)	6.79
		Cer (51:1)	6.80
Identification problems: misidentifications			
PC 37:7 | PC 15:1_22:6	6.00	PE (40:7)	6.00

Insufficient data: in some instances, the software may simply make an identification where there are insufficient data for a definitive identification. An example is the confusion of PCs and PEs, closely related lipids based on a glycerol backbone, two fatty acid chains, a phosphate group, and a choline or ethanolamine group, respectively. MS DIAL tended to identify these features as PCs, and Lipostar tended to identify the features as PEs. In the example in Table 2 at t_R 6.00 min, manual inspection of the MS^2 spectrum indicated that there was insufficient information to be definitive either way (shown in Supplementary Materials, Figure S2).

4. Discussion

The ongoing issues around reproducibility in biostatistics and bioinformatics are well-described [43,44]. These issues have a number of causes, such as insufficient documentation, inappropriate applications of hypothesis testing, or poor study design. These challenges also extend to metabolomic and lipidomic biomarker identifications, with consequences for reproducibility when developing diagnostic and prognostic panels [5,45,46]. The challenge can be further exacerbated by the issue shown in this work of inconsistent identifications being produced by different software platforms, even from the same spectral data. Here, agreement on lipid identifications for a bulk cell lysate between Lipostar and MS DIAL, two open-source lipidomics platforms in common use, was just 14.0% overall. In addition, in-built scores for identification confidence were, in our view, insufficient for the complexities of the issue at hand, warranting further steps to define the data. The headline differences are partly attributable to different underlying databases (MS DIAL partly uses LipidBlast [47], and LipoStar uses LipidMAPS [48]). Inconsistencies can be reduced through the exclusion of MS^1 identifications, careful manual curation, and experimental iteration. Nonetheless, the lack of consistency can still present a challenge to researchers.

One approach to dealing with potential problems in LC–MS analysis is outlier detection. The novel SVM regression with the LOOCV method described here successfully identified the major physicochemical properties governing elution order. This was achieved by using H count for acyl length, the CH relationship to identify saturation as a latent variable [49], and automatically identifying the hydrophobicity of lipid headgroups and their influence on t_R, for example, correctly ordering PC and DG headgroups [50]. Crucially, the algorithm can incorporate all these latent variables in its decision-making instead of relying on one variable for assessment; a comparison based solely on C count, for example, cannot provide significant information about t_R. Interestingly, whilst equivalent carbon number and its relationship with retention time has been proposed as a means of verifying t_R [51], the data-driven approach described here finds better performance from H count in identifying the latent variable (in this case, acyl length), with C count exhibiting non-confounding redundancy. For MS DIAL, outlier detection estimated the peak identifications as 19.9% internally inconsistent and the Lipostar peak identifications as 30.9% internally inconsistent. Taken in combination with the limited overlap between identifications of just 14%, these observations are strongly suggestive that a significant proportion of 'top hit' lipid identifications would pose reproducibility problems.

MS^1 spectra are generally deemed insufficient for lipid identification, especially for QTOF instruments, which in many cases have lower mass accuracy than Orbitrap mass spectrometers. Consequently, the best practice in lipidomics is to use only MS^2 identifications [6], but even here, the agreement was not perfect. Only 36.1% of MS^2 identifications could be matched between MS DIAL and LipoStar, and the outlier detection algorithm found that 5.3% of MS DIAL and 22.4% of Lipostar putative MS^2 identifications were internally inconsistent. This lack of consistency in MS^2 identifications—both internally and between platforms—was driven by co-elution issues and closely related lipids being difficult to distinguish from each other. Consistency was greatest for the major lipid classes, especially TGs and PCs, and lowest for minor lipid classes. This inconsistency between different software platforms for MS^2 identifications presents a clear challenge for the interpretation of results, especially for bioinformaticians and clinicians who are less familiar with LC–MS workflows. These results demonstrate that 'top hit' identifications by a single software platform should never be taken as a 'given' by users of lipidomics research. Based on these findings, software versions and settings are as critical to best practice in reporting as instrumental settings and yet are not given the same prominence in lipid research or in guidelines such as the Lipidomics Minimal Reporting Checklist [10].

Many existing strategies exist to deal with the problem of misidentifications. Visualization tools such as Kendrick mass defect plots can help spot outliers but do not take into account the full range of variables available to the statistical learning approach described here [52]. Other strategies include the use of internal standards, but such standards are expensive, often only offer a small number of standards per lipid class, and can be deployed in a more focused manner once appropriate targets have been identified at the discovery stage. In this case study, the Avanti EquiSPLASH® LIPIDOMIX® standard was used, but this includes only 14 deuterated lipids in the major lipid classes, for example, including just one DG. This can confirm overall LC–MS performance and identify the rough t_R range for a lipid class within a run but is of lesser use in identifying specific lipids—increasing the number of deuterated standards can be prohibitively expensive, especially in discovery work. Inevitably, inaccurate identifications at the discovery stage pose costs and challenges later on. Data-driven approaches either rely on multiple repeats of the experiment to provide separate training and test sets such as QSSR [53] or the use of specific libraries for samples or methods, for example, data on t_R values for human plasma lipids or on collision cross sections [54,55]. These library-driven approaches will not reflect the wide range of analytical columns, mobile phases, platforms, and sampling matrices, especially at the discovery stage.

These results also demonstrate that reporting software platform confidence 'scores' can be helpful but is insufficient for definitive identification (and, at worst, may produce

unwarranted confidence). In addition to these automated steps of outlier detection and reporting of confidence criteria, manual inspection is essential for all potential biomarkers of interest. For example, a manual inspection can check for biological consistency, an important step in lipid review, e.g., flagging the presence of plant-based lipids such as sulfolipids as incongruous in a human plasma sample [56]. It will also frequently be the only means of checking the validity of outlier observations, distinguishing closely related lipids such as PCs and PEs, and resolving problems with co-eluting lipid identifications. There may even be merit in analyzing data using more than one lipidomics software platform—as shown here, neither MS DIAL nor Lipostar alone showed the level of sensitivity or specificity for lipid identifications that would be required for reproducible identifications. Other solutions include extending chromatography run-time, which can help with the co-elution of lipids; running samples in negative mode, as well as positive mode, is also strongly recommended for improved feature annotation. While not an option for the ZenoToF 7600 instrument used in this case study, for some instruments, polarity switching can also be used to offer positive and negative modes within a single run [57]. As with increasing the number of deuterated standards, these solutions can involve cost and sensitivity trade-offs and often require different gradients or phases. As an example, ammonium acetate is better for negative mode, whereas ammonium formate was used here. For targeted work, naturally, such steps become more practical.

This case study only compares two lipidomics platforms, MS DIAL and Lipostar. Many other platforms are available, such as Progenesis QI or LipidSearch, and a more comprehensive exercise to benchmark the full range of platforms across multiple test datasets (covering different biological matrices and instrument methodologies) would have considerable value in highlighting the strengths and weaknesses of each. As previously noted, operating in negative mode as well as positive mode would improve consistency, as certain lipid classes, such as TGs, ionize well only in positive polarity, whilst negative polarity can produce better outputs for phospholipids. In addition, both MS DIAL and Lipostar can import different libraries, and harmonizing the libraries used would reduce differential identifications. These are all issues that could be further addressed in future work and investigations. Nonetheless, our emphasis here is to highlight that in untargeted lipidomics—especially when using default settings and libraries—the complexity of multiple theoretical lipid identifications is one of the major pitfalls for (new) investigators. Many lipid MS studies contain serious errors [58], and this emphasizes the importance of the additional curation steps outlined here.

5. Conclusions

In this work, we demonstrate the challenges for reproducibility derived from the choice of lipidomics software platform, an under-investigated source of inconsistencies when identifying lipid biomarkers of interest. This is an important issue, especially for bioinformaticians and clinicians (or indeed generalist readers) when using analytical LC–MS outputs. We also show that a data-driven workflow for outlier detection can learn the latent variables that govern the order of elution and t_R, but manual curation will still be required. This is especially the case where MS^2 data are challenged by co-elution issues or where lipid classes are similar. In-built software scoring and checks are helpful but, in our view, insufficient, necessitating additional quality control workflow steps. These are essential to reduce inconsistencies in identifications when different groups use different lipidomics platforms, to address problems with reproducibility and replicability for end users of LC–MS data, and to improve confidence in bioinformatics analyses using lipid biomarkers.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo14080461/s1, Table S1: Mass spectrometry settings; Table S2: MS Dial settings for the PANC1 data; Table S3: Lipostar settings for the PANC1 data; Figure S1: Fragmentation spectra for co-eluting lipids, t_R between 6.78 and 6.80; Figure S2: Fragmentation

spectra for conflicting identifications contain different indicators of co-eluting PE (neutral loss of 141 Da), PC (fragment ion of 184 Da) and even PS (neutral loss of 185 Da).

Author Contributions: Conceptualization, M.S.; data curation, J.v.G., K.S. and A.O.; formal analysis, J.v.G. and K.S.; funding acquisition, M.J.B.; investigation, J.v.G. and K.S.; methodology, L.A.G., N.G. and M.S.; project administration, M.J.B., N.G. and M.S.; resources, M.J.B. and N.G.; software, J.v.G., K.S., L.A.G., A.O. and M.S.; supervision, M.S.; validation, M.J.B. and A.O.; visualization, M.S.; writing—original draft, J.v.G. and K.S.; writing—review and editing, L.A.G., N.G. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Biotechnology and Biological Sciences Research Council, grant number BB/W019116/1, and by the Engineering and Physical Sciences Research Council, grant numbers EP/R031118/1, EP/X015491/1 and EP/X034933/1.

Data Availability Statement: The Python code in .ipynb notebook format, as well as the raw mass spectrometry files, are available at the Zotero repository with the following URL: https://zenodo.org/records/10889321 (accessed on 28 March 2024); DOI: 10.5281/zenodo.10889320.

Acknowledgments: The authors wish to acknowledge the work of the Metabolomics Standards Initiative and the Lipidomics Standards Initiative.

Conflicts of Interest: Author Lee A. Gethings was employed by the company Waters Corporation (Wilmslow, UK). The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wenk, M.R. The Emerging Field of Lipidomics. *Nat. Rev. Drug Discov.* **2005**, *4*, 594–610. [CrossRef] [PubMed]
2. Han, X. Lipidomics for Studying Metabolism. *Nat. Rev. Endocrinol.* **2016**, *12*, 668–679. [CrossRef] [PubMed]
3. Hasin, Y.; Seldin, M.; Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol.* **2017**, *18*, 83. [CrossRef]
4. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-Omics Data Integration, Interpretation, and Its Application. *Bioinform Biol. Insights* **2020**, *14*, 117793221989905. [CrossRef] [PubMed]
5. Sarmad, S.; Viant, M.R.; Dunn, W.B.; Goodacre, R.; Wilson, I.D.; Chappell, K.E.; Griffin, J.L.; O'Donnell, V.B.; Naicker, B.; Lewis, M.R.; et al. A Proposed Framework to Evaluate the Quality and Reliability of Targeted Metabolomics Assays from the UK Consortium on Metabolic Phenotyping (MAP/UK). *Nat. Protoc.* **2023**, *18*, 1017–1027. [CrossRef] [PubMed]
6. Köfeler, H.C.; Ahrends, R.; Baker, E.S.; Ekroos, K.; Han, X.; Hoffmann, N.; Holčapek, M.; Wenk, M.R.; Liebisch, G. Recommendations for Good Practice in MS-Based Lipidomics. *J. Lipid Res.* **2021**, *62*, 100138. [CrossRef] [PubMed]
7. Theodoridis, G.; Gika, H.; Raftery, D.; Goodacre, R.; Plumb, R.S.; Wilson, I.D. Ensuring Fact-Based Metabolite Identification in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **2023**, *95*, 3909–3916. [CrossRef] [PubMed]
8. Lipidomics Standards Initiative Consortium. Lipidomics Needs More Standardization. *Nat. Metab.* **2019**, *1*, 745–747. [CrossRef]
9. Lipidomics Standards Initiative. Available online: https://lipidomicstandards.org/ (accessed on 20 June 2023).
10. McDonald, J.G.; Ejsing, C.S.; Kopczynski, D.; Holčapek, M.; Aoki, J.; Arita, M.; Arita, M.; Baker, E.S.; Bertrand-Michel, J.; Bowden, J.A.; et al. Introducing the Lipidomics Minimal Reporting Checklist. *Nat. Metab.* **2022**, *4*, 1086–1088. [CrossRef]
11. MSI Board Members; Sansone, S.-A.; Fan, T.; Goodacre, R.; Griffin, J.L.; Hardy, N.W.; Kaddurah-Daouk, R.; Kristal, B.S.; Lindon, J.; Mendes, P.; et al. The Metabolomics Standards Initiative. *Nat. Biotechnol.* **2007**, *25*, 846–848. [CrossRef]
12. Spicer, R.A.; Salek, R.; Steinbeck, C. A Decade after the Metabolomics Standards Initiative It's Time for a Revision. *Sci. Data* **2017**, *4*, 170138. [CrossRef]
13. Saunders, K.D.G.; Von Gerichten, J.; Lewis, H.-M.; Gupta, P.; Spick, M.; Costa, C.; Velliou, E.; Bailey, M.J. Single-Cell Lipidomics Using Analytical Flow LC-MS Characterizes the Response to Chemotherapy in Cultured Pancreatic Cancer Cells. *Anal. Chem.* **2023**, *95*, 14727–14735. [CrossRef] [PubMed]
14. Avela, H.F.; Sirén, H. Advances in Lipidomics. *Clin. Chim. Acta* **2020**, *510*, 123–141. [CrossRef]
15. Song, H.; Ladenson, J.; Turk, J. Algorithms for Automatic Processing of Data from Mass Spectrometric Analyses of Lipids. *J. Chromatogr. B* **2009**, *877*, 2847–2854. [CrossRef] [PubMed]
16. Kanu, A.B.; Dwivedi, P.; Tam, M.; Matz, L.; Hill, H.H. Ion Mobility-Mass Spectrometry. *J. Mass Spectrom.* **2008**, *43*, 1–22. [CrossRef] [PubMed]
17. Fedorova, E.S.; Matyushin, D.D.; Plyushchenko, I.V.; Stavrianidi, A.N.; Buryak, A.K. Deep Learning for Retention Time Prediction in Reversed-Phase Liquid Chromatography. *J. Chromatogr. A* **2022**, *1664*, 462792. [CrossRef]
18. Habra, H.; Meijer, J.L.; Shen, T.; Fiehn, O.; Gaul, D.A.; Fernández, F.M.; Rempfert, K.R.; Metz, T.O.; Peterson, K.E.; Evans, C.R.; et al. metabCombiner 2.0: Disparate Multi-Dataset Feature Alignment for LC-MS Metabolomics. *Metabolites* **2024**, *14*, 125. [CrossRef]
19. Krokhin, O.V.; Spicer, V. Predicting Peptide Retention Times for Proteomics. *Curr. Protoc. Bioinform.* **2010**, *13*, 13–14. [CrossRef]

20. Baczek, T.; Kaliszan, R. Predictions of Peptides' Retention Times in Reversed-Phase Liquid Chromatography as a New Supportive Tool to Improve Protein Identification in Proteomics. *Proteomics* **2009**, *9*, 835–847. [CrossRef]
21. Henneman, A.; Palmblad, M. Retention Time Prediction and Protein Identification. *Methods Mol. Biol.* **2020**, *2051*, 115–132. [CrossRef]
22. Pfeifer, N.; Leinenbach, A.; Huber, C.G.; Kohlbacher, O. Statistical Learning of Peptide Retention Behavior in Chromatographic Separations: A New Kernel-Based Approach for Computational Proteomics. *BMC Bioinform.* **2007**, *8*, 468. [CrossRef] [PubMed]
23. Kell, P.; Sidhu, R.; Qian, M.; Mishra, S.; Nicoli, E.-R.; D'Souza, P.; Tifft, C.J.; Gross, A.L.; Gray-Edwards, H.L.; Martin, D.R.; et al. A Pentasaccharide for Monitoring Pharmacodynamic Response to Gene Therapy in GM1 Gangliosidosis. *eBioMedicine* **2023**, *92*, 104627. [CrossRef]
24. Field, A.P.; Gillett, R. How to Do a Meta-analysis. *Brit. J. Math. Statis* **2010**, *63*, 665–694. [CrossRef]
25. O'Donnell, V.B.; Ekroos, K.; Liebisch, G.; Wakelam, M. Lipidomics: Current State of the Art in a Fast Moving Field. *WIREs Mech. Dis.* **2020**, *12*, e1466. [CrossRef]
26. Zhang, H.; Gao, Y.; Sun, J.; Fan, S.; Yao, X.; Ran, X.; Zheng, C.; Huang, M.; Bi, H. Optimization of Lipid Extraction and Analytical Protocols for UHPLC-ESI-HRMS-Based Lipidomic Analysis of Adherent Mammalian Cancer Cells. *Anal. Bioanal. Chem.* **2017**, *409*, 5349–5358. [CrossRef]
27. Cajka, T.; Smilowitz, J.T.; Fiehn, O. Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography–High-Resolution Mass Spectrometry Platforms. *Anal. Chem.* **2017**, *89*, 12360–12368. [CrossRef] [PubMed]
28. Goracci, L.; Tortorella, S.; Tiberi, P.; Pellegrino, R.M.; Di Veroli, A.; Valeri, A.; Cruciani, G. Lipostar, a Comprehensive Platform-Neutral Cheminformatics Tool for Lipidomics. *Anal. Chem.* **2017**, *89*, 6257–6264. [CrossRef]
29. Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; et al. A Lipidome Atlas in MS-DIAL 4. *Nat. Biotechnol.* **2020**, *38*, 1159–1163. [CrossRef] [PubMed]
30. Waldmann, E. Quantile Regression: A Short Story on How and Why. *Stat. Model.* **2018**, *18*, 203–218. [CrossRef]
31. Koenker, R. Quantile Regression: 40 Years On. *Annu. Rev. Econ.* **2017**, *9*, 155–176. [CrossRef]
32. Gottard, A.; Vannucci, G.; Grilli, L.; Rampichini, C. Mixed-Effect Models with Trees. *Adv. Data Anal. Classif.* **2023**, *17*, 431–461. [CrossRef]
33. Alkharusi, H. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *Int. J. Educ.* **2012**, *4*, 202. [CrossRef]
34. Lundberg, S.M.; Erion, G.G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2018**, arXiv:1802.03888. [CrossRef]
35. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Long Beach, CA, USA, 2017; Volume 30.
36. Covert, I.; Lundberg, S.; Lee, S.-I. Feature Removal Is a Unifying Principle for Model Explanation Methods. *arXiv* **2020**, arXiv:2011.03623. [CrossRef]
37. Boyer, G. Chemparse 2022. Available online: https://pypi.org/project/chemparse/ (accessed on 25 October 2023).
38. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
39. Release Notes—SHAP Latest Documentation. Available online: https://shap.readthedocs.io/en/latest/release_notes.html (accessed on 25 October 2023).
40. Da Costa-Luis, C.; Larroque, S.K.; Altendorf, K.; Mary, H.; Richardsheridan; Korobov, M.; Yorav-Raphael, N.; Ivanov, I.; Bargull, M.; Rodrigues, N.; et al. Tqdm: A Fast, Extensible Progress Bar for Python and CLI 2023. Available online: https://github.com/tqdm/tqdm (accessed on 25 October 2023).
41. Raybaut, P. Spyder IDE. Available online: https://www.spyder-ide.org/ (accessed on 20 June 2023).
42. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. *Nat. Methods* **2015**, *12*, 523–526. [CrossRef] [PubMed]
43. Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*, e124. [CrossRef] [PubMed]
44. Leek, J.T.; Jager, L.R. Is Most Published Research Really False? *Annu. Rev. Stat. Appl.* **2017**, *4*, 109–122. [CrossRef]
45. Wood, P.L.; Cebak, J.E. Lipidomics Biomarker Studies: Errors, Limitations, and the Future. *Biochem. Biophys. Res. Commun.* **2018**, *504*, 569–575. [CrossRef]
46. Onoja, A.; Von Gerichten, J.; Lewis, H.-M.; Bailey, M.J.; Skene, D.J.; Geifman, N.; Spick, M. Meta-Analysis of COVID-19 Metabolomics Identifies Variations in Robustness of Biomarkers. *Int. J. Mol. Sci.* **2023**, *24*, 14371. [CrossRef]
47. Kind, T.; Liu, K.-H.; Lee, D.Y.; DeFelice, B.; Meissen, J.K.; Fiehn, O. LipidBlast in Silico Tandem Mass Spectrometry Database for Lipid Identification. *Nat. Methods* **2013**, *10*, 755–758. [CrossRef]
48. Conroy, M.J.; Andrews, R.M.; Andrews, S.; Cockayne, L.; Dennis, E.A.; Fahy, E.; Gaud, C.; Griffiths, W.J.; Jukes, G.; Kolchin, M.; et al. LIPID MAPS: Update to Databases and Tools for the Lipidomics Community. *Nucleic Acids Res.* **2024**, *52*, D1677–D1682. [CrossRef] [PubMed]
49. Ovčačíková, M.; Lísa, M.; Cífková, E.; Holčapek, M. Retention Behavior of Lipids in Reversed-Phase Ultrahigh-Performance Liquid Chromatography-Electrospray Ionization Mass Spectrometry. *J. Chromatogr. A* **2016**, *1450*, 76–85. [CrossRef] [PubMed]

50. Pchelkin, V.P. Calculations of the Hydrophobicity of Lipid Molecules by the Elution Strength of the Chromatographic Solvent. *J. Anal. Chem.* **2020**, *75*, 615–621. [CrossRef]
51. White, J.B.; Trim, P.J.; Salagaras, T.; Long, A.; Psaltis, P.J.; Verjans, J.W.; Snel, M.F. Equivalent Carbon Number and Interclass Retention Time Conversion Enhance Lipid Identification in Untargeted Clinical Lipidomics. *Anal. Chem.* **2022**, *94*, 3476–3484. [CrossRef]
52. 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.* **2001**, *73*, 4676–4681. [CrossRef]
53. Naylor, B.C.; Catrow, J.L.; Maschek, J.A.; Cox, J.E. QSRR Automator: A Tool for Automating Retention Time Prediction in Lipidomics and Metabolomics. *Metabolites* **2020**, *10*, 237. [CrossRef]
54. Vu, N.; Narvaez-Rivas, M.; Chen, G.-Y.; Rewers, M.J.; Zhang, Q. Accurate Mass and Retention Time Library of Serum Lipids for Type 1 Diabetes Research. *Anal. Bioanal. Chem.* **2019**, *411*, 5937–5949. [CrossRef] [PubMed]
55. Rose, B.S.; May, J.C.; Picache, J.A.; Codreanu, S.G.; Sherrod, S.D.; McLean, J.A. Improving Confidence in Lipidomic Annotations by Incorporating Empirical Ion Mobility Regression Analysis and Chemical Class Prediction. *Bioinformatics* **2022**, *38*, 2872–2879. [CrossRef]
56. Shimojima, M. Biosynthesis and Functions of the Plant Sulfolipid. *Prog. Lipid Res.* **2011**, *50*, 234–239. [CrossRef] [PubMed]
57. Abou-Elwafa Abdallah, M.; Nguyen, K.-H.; Ebele, A.J.; Atia, N.N.; Ali, H.R.H.; Harrad, S. A Single Run, Rapid Polarity Switching Method for Determination of 30 Pharmaceuticals and Personal Care Products in Waste Water Using Q-Exactive Orbitrap High Resolution Accurate Mass Spectrometry. *J. Chromatogr. A* **2019**, *1588*, 68–76. [CrossRef] [PubMed]
58. Skotland, T.; Ekroos, K.; McDonald, J.; Ahrends, R.; Liebisch, G.; Sandvig, K. Pitfalls in Lipid Mass Spectrometry of Mammalian Samples—A Brief Guide for Biologists. *Nat. Rev. Mol. Cell Biol.* **2024**, *1471*, 80. [CrossRef]