analytical chemistry

Article

pubs.acs.org/ac

# Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics

Justin J. J. van der Hooft,[*,†,‡] Joe Wandy,[†] Francesca Young,[†] Sandosh Padmanabhan,[‡] Konstantinos Gerasimidis,[§] Karl E. V. Burgess,[†] Michael P. Barrett,[†,∥] and Simon Rogers[*,†,⊥]

[†]Glasgow Polyomics, University of Glasgow, Glasgow G61 1HQ, United Kingdom

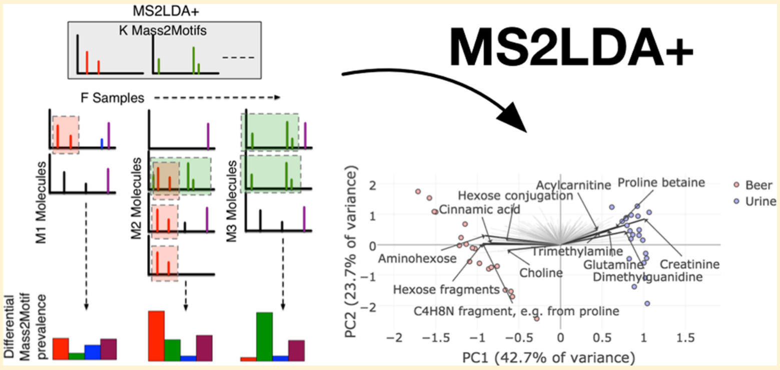
[‡]Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom

[§]Human Nutrition, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, New Lister Building, Glasgow Royal Infirmary, Glasgow G31 2ER, United Kingdom

[∥]Wellcome Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow G12 8TA, United Kingdom

[⊥]School of Computing Science, University of Glasgow, Glasgow G12 8RZ, United Kingdom

*Supporting Information*



**ABSTRACT:** In untargeted metabolomics approaches, the inability to structurally annotate relevant features and map them to biochemical pathways is hampering the full exploitation of many metabolomics experiments. Furthermore, variable metabolic content across samples result in sparse feature matrices that are statistically hard to handle. Here, we introduce MS2LDA+ that tackles both above-mentioned problems. Previously, we presented MS2LDA, which extracts biochemically relevant molecular substructures ("Mass2Motifs") from a collection of fragmentation spectra as sets of co-occurring molecular fragments and neutral losses, thereby recognizing building blocks of metabolomics. Here, we extend MS2LDA to handle multiple metabolomics experiments in one analysis, resulting in MS2LDA+. By linking Mass2Motifs across samples, we expose the variability in prevalence of structurally related metabolite families. We validate the differential prevalence of substructures between two distinct samples groups and apply it to fecal samples. Subsequently, within one sample group of urines, we rank the Mass2Motifs based on their variance to assess whether xenobiotic-derived substructures are among the most-variant Mass2Motifs. Indeed, we could ascribe 22 out of the 30 most-variant Mass2Motifs to xenobiotic-derived substructures including paracetamol/acetaminophen mercapturate and dimethylpyrogallol. In total, we structurally characterized 101 Mass2Motifs with biochemically or chemically relevant substructures. Finally, we combined the discovered metabolite families with full scan feature intensity information to obtain insight into core metabolites present in most samples and rare metabolites present in small subsets now linked through their common substructures. We conclude that by biochemical grouping of metabolites across samples MS2LDA+ aids in structural annotation of metabolites and guides prioritization of analysis by using Mass2Motif prevalence.

Large high-throughput metabolomics experiments are becoming more prevalent across many areas of medicine and the life sciences.[1,2] Analysis of the resulting data sets is challenging and current techniques fail to extract all of the rich structure they encapsulate. Most techniques work at the level of individual mass features (molecules); finding those that change systematically across experimental groups before attempting to identify them. Although analyses sometimes map the identified molecules to molecular networks,[2,3] relationships that exist between molecules are rarely used earlier in the analysis.

**Received:** April 13, 2017
**Accepted:** June 16, 2017
**Published:** June 16, 2017

DOI: 10.1021/acs.analchem.7b01391

Identification of mass features (mapping them to molecular structures) is the main bottleneck in untargeted metabolomics. It is commonly recognized that gas-phase mass fragmentation experiments are essential to support metabolite annotations.[4,5] The resulting fragmentation data is complex and using it to perform annotation and identification is challenging.[4,5] Comparing measured fragment spectra with spectral databases is largely ineffective (a recent commentary estimated that on average around 2% of molecules in a typical experiment could be confidently identified in this way[5]). Moreover, although databases will continue to grow, they can only be populated with "known unknowns" (i.e., previously studied molecules) hindering discovery of novel natural products or unexpected metabolites. Focusing on individual mass features also requires chromatographic retention time alignment across the samples which becomes increasingly challenging as the number of samples increases and effectively precludes comparisons between different chromatographic platforms. Retention time alignment makes it hard to deal with molecules that only appear in a small subset of the samples as they will often be considered to be noise and removed. However, in some studies, molecules of interest may appear in only a small subset of the samples (e.g., a drug or other xenobiotic).

One strategy to overcome these limitations is to make better use of fragmentation data. Recently, multiple tools have been proposed for processing, analysis, and visualization of fragmentation data sets.[6−10] Of these, MS2Analyzer[6] relies on knowledge of the biochemistry of interest and fragment/loss patterns of importance, which is only of limited use in experiments where the goal is to *uncover* unknown biochemistry. Molecular Networking[7,11] groups spectra according to their overall level of similarity. Those spectra that can be structurally identified can annotate their neighbors, propagating metabolite annotations to previously unknown molecules.[5] However, each spectrum can only belong to one group even though many molecules consist of multiple biochemically diverse substructures. Should adenosine, which comprises adenine and pentose (ribose) substructures, be grouped with other adenine containing molecules or ribose containing ones? MetFamily[8] was recently introduced to sidestep the identification problem by clustering MS2 spectra to find *structural families* and linking these clusters to differential expression, thereby revealing regulated metabolite families. While this approach neatly integrates MS1 and MS2 data, the cluster analysis does not have the flexibility to represent molecules consisting of multiple diverse substructures.

Recently, we developed MS2LDA[10] for exploration of fragment data. On the basis of Latent Dirichlet Allocation (LDA),[12] MS2LDA decomposes each molecule into one or more Mass2Motifs, allowing for more efficient molecular grouping, searching, and exploration. Mass2Motifs consist of fragments and losses conserved across multiple spectra and often correspond to chemical substructures. Spectra sharing a Mass2Motif can be grouped even if the Mass2Motif only accounts for a small portion of their spectra and, as they consist of multiple Mass2Motifs, spectra can belong to multiple structural families. We have previously shown that MS2LDA can decompose metabolites into biochemically relevant substructures, such as amino acid, nucleotide, or hexose related Mass2Motifs.[10]

MS2LDA exploits biochemical similarities within a single fragmentation run (i.e., one DDA data set) and therefore does not tackle a key step in many analyses, the direct highlighting and extraction of changes in metabolomes across samples. Here, we introduce MS2LDA+, an extension of MS2LDA that simultaneously analyzes multiple samples, across which the Mass2Motifs are shared. Crucially, this sharing means we can measure the change in prevalence of Mass2Motifs across samples allowing us to perform differential analysis of Mass2Motifs (something that was not possible with MS2LDA). The development of this new model is motivated by the observation that very similar Mass2Motifs were found independently (through manual comparison) across multiple samples.[10] MS2LDA+ formalizes and automates this laborious matching process by processing multiple samples with the same set of Mass2Motifs, allowing us to decompose molecules from different samples into a shared set of substructures and automatically compare the *prevalence* of Mass2Motifs across the samples.

To validate our MS2LDA+ model, we apply it to a data set consisting of 19 beer samples and 22 urine samples, hypothesizing that MS2LDA+ will discover substructures with differential prevalence (i.e., substructures that are beer and urine specific). The MS2LDA+ pipeline is then applied to perform fragmentome-based molecular phenotyping of 22 urine samples from a cohort of stroke patients and finally applied to a set of fecal samples obtained from children with Crohn's disease at different time points during nutritional therapy. Previously, we have shown that untargeted mass spectrometry fragmentation experiments can expose not only the presence of different classes of antihypertensive drugs and their metabolites in urine samples but also over-the-counter drugs and endogenous metabolite families.[13] By applying the MS2LDA+ pipeline to discover fragmentation patterns across urine samples, we aim to structurally characterize the discovered urinary metabolite families. Our hypothesis is that Mass2Motifs with highly variable prevalence, representing drug (or other xenobiotic) substructures, appear in only a subset of samples, whereas Mass2Motifs that display low variance represent common endogenous substructures, such as acylcarnitines and acylglutamines.[13] This paper introduces the concept of MS2LDA+ that allows for the determination of Mass2Motif's prevalence across multiple samples.

## ■ MATERIAL AND METHODS

**Materials.** *Urine Samples.* Urine samples from anonymized human volunteers were used from a clinical sample set in the Glasgow Polyomics' archive. These samples were obtained as part of a trial for which ethical approval was applied for through the Multi-Centre Research and Ethics Committee (MREC), which was granted by the Scottish MREC and (with MREC No. 06/MRE00/106). Informed consent was obtained from all individual study participants. Spot urine samples were obtained from the cohort of elderly patients upon their first admission in the clinic. A different subset as in[13] was chosen: urine extracts of 22 patients were selected as follows, diagnosed with stroke, administering a variety of drugs including a number of antihypertensives, and availability of the sample extract in the Glasgow Polyomics' archive. The resulting subject's age range spanned from 52 to 85; 13 were male, and 9 female. More details can be found in Supporting Information section S1 and Table S1.

*Beer Samples.* The 10 mL samples of 18 different beers were collected from bottles over a period of 5 months. One beer was sampled twice from different bottles. Details can be found in the Supporting Information section S2.

*Stool Samples.* Stool samples originated from two children with active Crohn's disease (9.2 and 12.9 years) who received disease induction treatment with exclusive enteral nutrition (EEN) as described previously.[14,15] Both children entered clinical remission and their fecal calprotectin, a marker of colonic inflammation decreased significantly at the end of their treatment. In total, five serial stool samples were collected per patient and a single one from two healthy controls (10.7 and 11.2 years). From CD children, a first sample was collected before EEN, three samples were collected during EEN (at ∼15, 30, and 56 days), and a final sample was collected when patients returned to their habitual diet (∼60 days after EEN cessation). Stool samples were collected within 2 h of defecation, homogenized with mechanical kneading immediately, and aliquots were stored at −80 °C until further analysis. Carers and participants provided written informed consent, and the study was approved by the local research ethics committee (Reference Number 05/S0708/66).

*Chemicals.* HPLC-grade methanol, acetonitrile, isopropanol, and analytical reagent grade chloroform were acquired from Fisher Scientific, Loughborough, U.K. HPLC grade $H_2O$ was purchased from VWR Chemicals, Fountenay-sous-Bois, France. Formic acid (for mass spectrometry) and ammonium carbonate were acquired from Fluka Analytical (Sigma-Aldrich), Steinheim, Germany.

## METHODS

**Sample Preparation.** A general metabolome extraction procedure was performed:[16] (i) 5 $\mu$L urine was extracted in 200 $\mu$L of chloroform/methanol/water (1:3:1) at 4 °C; (ii) then vortexed for 5 min at 4 °C; (iii) then centrifuged for 3 min (13 000*g*) at 4 °C. The resulting supernatant was stored at −80 °C until analysis. A pooled aliquot of the 22 selected urine samples was prepared prior to the LC−MS runs with DDA applying higher collision dissociation (HCD) as in ref 13. The same procedure was followed for the 19 beer samples, see also ref 10. Finally, to create fecal extracts, the stool samples were freeze-dried and 5 mg of lyophilized fecal material was extracted in 200 $\mu$L of chloroform/methanol/water (1:3:1) at 4 °C; followed by homogenization in a FastPrep-24 homogenizer for 60 s at stroke setting 5, after which the same procedure as for urine was followed.

**Analytical Platform.** A Thermo Scientific Ultimate 3000 RSLCnano liquid chromatography system (Thermo Scientific, CA) was used. That system was coupled to a Thermo Scientific Q-Exactive Orbitrap mass spectrometer equipped with a HESI II interface (Thermo Scientific, Hemel Hempstead, U.K.). Thermo Xcalibur Tune software (version 2.5) was used for instrument control and data acquisition.

**LC Settings.** The HILIC separation was performed with a SeQuant ZIC-pHILIC column (150 mm × 4.6 mm, 5 $\mu$m) equipped with the corresponding precolumn (Merck KGaA, Darmstadt, Germany). A linear biphasic LC gradient was conducted from 80% B to 20% B over 15 min, followed by a 2 min wash with 5% B, and 7 min re-equilibration with 80% B, where solvent B is acetonitrile and solvent A is 20 mM ammonium carbonate in water. The flow rate was 300 $\mu$L/min, column temperature was maintained at 25 °C, injection volume was 10 $\mu$L, and samples were maintained at 4 °C in the autosampler.[13]

**MS and MS/MS Settings.** MS and MS/MS settings used to generate separate mode fragmentation files are fully described in refs 10 and 13. In short, for positive- and negative-ionization separate fragmentation modes, the duty cycles consisted of a full scan in positive-ionization mode, followed by a TopN data-dependent MS/MS (MS2) fragmentation event taking the 10 most abundant ion species not on the dynamic exclusion list. MS/MS fragmentation spectra were acquired using stepped higher collision dissociation combining 25.2, 60.0, and 94.8 normalized collision energies in one MS2 scan. In full-scan mode, the duty cycle consisted of two full-scan events alternating positive and negative ionization modes.

## DATA ACQUISITION AND PROCESSING

**Data Acquisition.** Quality control procedures from Glasgow Polyomics were used.[16] Details can be found in the Supporting Information section S3.

**Data Processing: Feature Extraction.** Data, in the form of .mzXML (full scan) and .mzML (fragmentation) files, are preprocessed using XCMS[17] and MzMatch[18] for peak detection and RMassBank[19] for detecting MS1−MS2 pairs, before matrix formation by aligning MS2 features across different spectra and samples. One benefit of working at the level of the structural families defined by MS2LDA is that the samples can be preprocessed separately and no retention time alignment is required. The resulting data set consists of a set of MS2 spectra for each sample, containing fragment and loss features (and their intensities) that have been matched across all samples (Supporting Information section S3).

**MS2LDA+ Model.** The MS2LDA+ model is an extension to standard LDA in which a single set of motifs (known as topics in standard LDA) is shared across multiple samples. For inference, we have developed a Variational Bayes[12] scheme in which each sample is modeled with standard MS2LDA except for the updates of the motifs which are pooled across the samples (Supporting Information section S4). The output is a set of Mass2Motifs and assignments of Mass2Motifs to each MS1 peak in each sample. Essentially, all mass fragments and neutral losses are now linked across samples to assess their co-occurrence in the entire collection of fragmentation spectra while also storing feature matrices for each individual sample. The extension from MS2LDA to MS2LDA+ model is depicted in Figure 1.

The MS2LDA pipeline was extended to MS2LDA+ to enable the analysis of multiple samples in one analysis by linking all fragments and losses and their co-occurrences across the entire corpus while storing sample-specific information (Figure 1). In MS2LDA+, the sample-specific prevalence of the different Mass2Motifs for the $f$th sample is captured by a parameter vector $\alpha^f$, that has one value per Mass2Motif ($\alpha_k^f$) (represented by the bar charts at the bottom of Figure 1). The higher the value, the more prevalent that Mass2Motif is within the sample. Prevalence here can be interpreted as the proportion of all feature intensity (of the fragmented molecules) that is explained by this Mass2Motif. These $\alpha$ vectors provide a high-level view of the biochemical makeup of each sample in terms of the prevalence of the different, shared Mass2Motifs (Figure 1).

Following MS2LDA, we used the Mass2Motif-molecule probabilities to define the links between the molecules. The probability can be interpreted as the proportion of a molecule's spectrum that is explained by this Mass2Motif, and it is therefore affected by the number of peaks in the molecule's spectrum. For example, two molecules that both include a complete Mass2Motif would have very different probabilities if one of them had many more other peaks. In a further

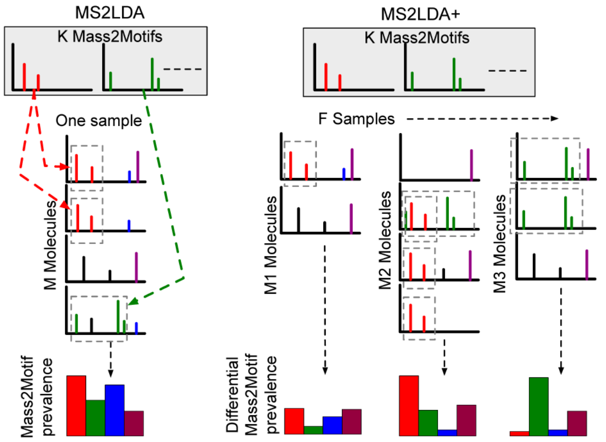


**Figure 1.** Extension from MS2LDA to MS2LDA+. In MS2LDA, a single sample (containing *M* molecules) is decomposed using *K* Mass2Motifs. In MS2LDA+, *F* samples (containing M1, M2, M3 molecules, etc.) are decomposed onto shared Mass2Motifs. Prevalence of the Mass2Motifs can then be compared across the samples. In this example, the red Mass2Motif is most prevalent in the second sample, and the green in the third.

development from the original MS2LDA, we have therefore implemented a second, complementary linking score: the *overlap score*. The overlap score measures how much of the Mass2Motif is present in the spectrum rather than how much of the spectrum is explained by the Mass2Motif (Supporting Information section S5). The higher that score, the more mass fragments and neutral losses from the Mass2Motif can be found in the fragmentation spectrum of the parent ion, increasing the confidence that the substructure is indeed present. An intrinsic model property is that all spectra must be assigned to at least one Mass2Motif. This enforces "alien" metabolites not sharing any substructures with other metabolites to be part of a structural family (with very high probabilities) despite sharing hardly any or no characteristic features. These outliers are now easily recognized by the combination of a high membership probability but very low overlap score.

## DATA ANALYSIS

**Statistical Tools.** *PCA Analysis.* PCA analysis was performed on the 41 × 500 matrix of $\alpha$ values to project it down into a 41 × 2 matrix for visualization. The values were normalized so that the total value within each sample was equal to 1. This allows the values to be interpreted as Mass2Motif probabilities within each sample. The standard approach of whitening the variables (motifs) prior to PCA was performed. PCA analysis was done in Python using the PCA method provided in the scikit-learn package.[20]

Mass2Motif differential prevalence was determined by computing the z-score for each Mass2Motif between the two groups (difference of $\alpha$ means divided by the sum of the standard deviations). As for PCA, the $\alpha$ values within each sample were normalized to sum to 1. To compute similarity between Mass2Motifs, their $\alpha$ values (now the vector of values across samples for each Mass2Motif) were compared by computing their pairwise Pearson correlation values.

**Structural Characterization of Mass2Motifs.** Mass2Motifs discovered in beer and urine were structurally characterized by comparison to earlier discovered Mass2Motifs[10] and characteristic fragments found by manual inspection of clusters in a Molecular Network,[13] through expert knowledge and matching of the Mass2Motif spectra to reference spectra in MzCloud (www.mzcloud.org).

**Data Availability.** All data, processed data, and codes used for this paper will be available for download from the university repository (http://researchdata.gla.ac.uk/402/). In addition, data is available through GNPS/MassIVE: MassIVE data set MSV000081118 contains the urine sample data, MSV000081119 the beer sample data, and MSV000081120 the stool (fecal) sample data. All codes can be found in GitHub (https://github.com/sdrogers/lda).

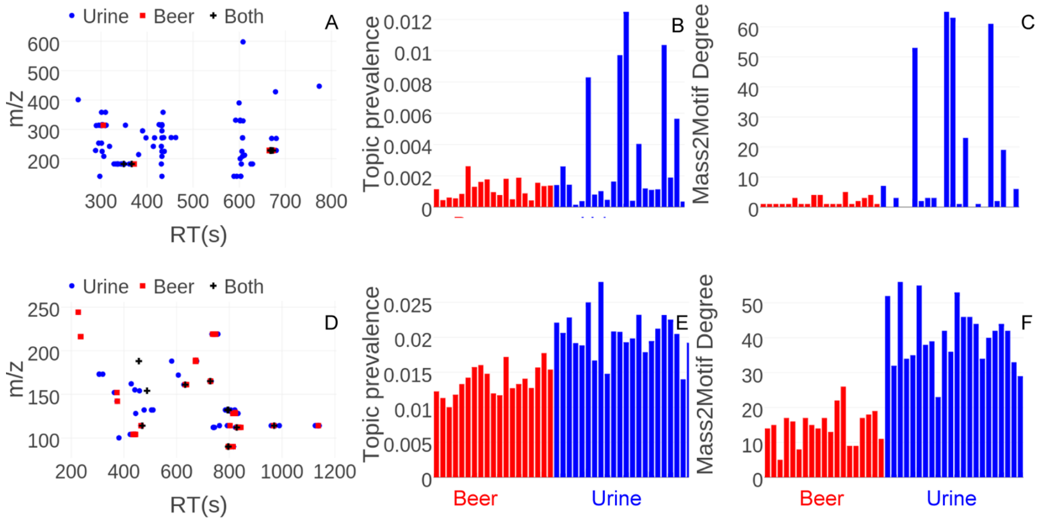


**Figure 2.** (A–C) Paracetamol mercapturate Mass2Motif in beer and urine samples with (A) metabolites displayed in $m/z$ vs RT plot, (B) $\alpha^f$, and (C) degrees. Note that the model finds (almost) no molecules from beer (as expected) that contain the paracetamol mercapturate Mass2Motif; those that do spuriously match are doing so because one or two abundant fragments overlap but the characteristic paracetamol mercapturate pattern is clearly absent from those beer fragmentation spectra. (D–F) Acetyl loss Mass2Motif in beer and urine samples with (D) metabolites displayed in $m/z$ vs RT plot, (E) $\alpha^f$, and (F) degrees. Note that we observe many molecules from beer, many from urine, and many that appear in both that contain the acetyl loss Mass2Motif.

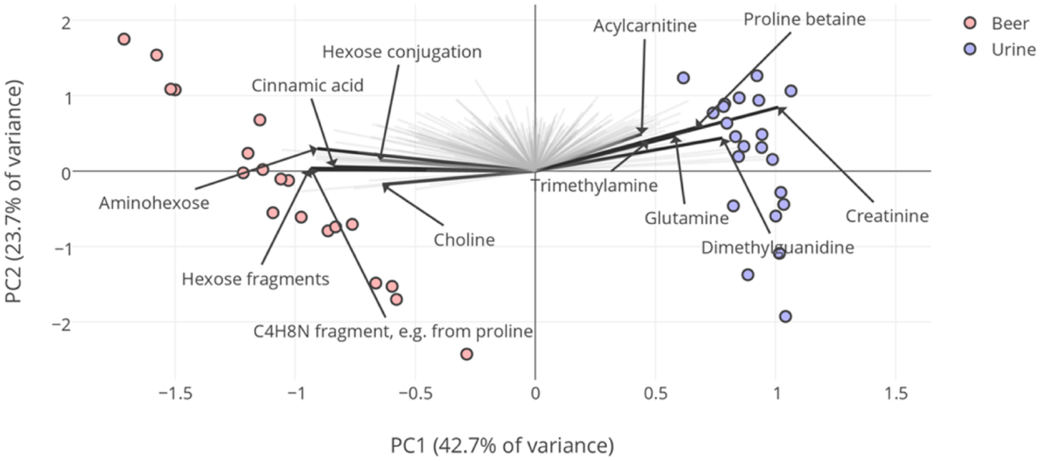


**Figure 3.** Principal component analysis (PCA) of 19 beer and 22 urine samples. The largest variance (42.7%) is explained by the differences between the beer and urine groups. Motifs are indicated with lines and several structurally characterized Mass2Motifs are highlighted.

## ■ RESULTS

**MS2LDA+ Exposes Biochemical Differences between Samples.** To validate the extent to which MS2LDA+ generates biochemically relevant knowledge from untargeted metabolomics experiments and the extent to which $\alpha^f$ provides a biochemical *summary* of each sample, sets of 19 beer and 22 urine samples were combined in one analysis to discover 300 Mass2Motifs. Using biological samples from two distinct origins will highlight the high-level differences that MS2LDA+ can extract from large experiments. The prevalence of two discovered Mass2Motifs across all the 41 samples is displayed in Figure 2. Figure 2A−C displays the paracetamol mercapturate with Figure 2A displaying the fragmented ions that contain the paracetamol mercapturate Mass2Motif, Figure 2B displaying the $\alpha^f$ across all samples, and Figure 2C showing the number of metabolites that passed a threshold of 0.01 on their molecule-Mass2Motif probability. As can be seen, the trend of the two histograms is very similar. The Mass2Motif is clearly prevalent in 5 of the urine samples and present in a couple more and virtually absent in the beer samples (those that are grouped from beer are incidentally mass-matched metabolites sharing only one or two abundant mass fragments), as expected since it is a human drug related Mass2Motif. Inspection revealed that the characteristic fragmentation pattern of paracetamol mercapturate consisting of 5 recurring fragments is absent from those beer fragmentation spectra. Higher threshold values on document-Mass2Motif probabilities and overlap scores would reduce the number of such spurious hits. The acetyl loss Mass2Motif (Figure 2D−F) displays a completely different picture of a fairly constant number of metabolites that contain this Mass2Motif in both beer and urine samples with some of them present in both sample sets (see Figure 2D).

**Substructure-Based Principal Component Analysis.** To visualize the results at a higher level, principal component analysis (PCA) was performed to project the 41 (total number of samples) $\alpha^f$ vectors from their original 300-dimensional space (one dimension per Mass2Motif) into two dimensions. Figure 3 shows clear separation of the two sample groups with the Mass2Motif loadings indicated as gray lines, highlighting the ability of MS2LDA+ to characterize complex mixtures. We subsequently explored the Mass2Motifs with high loadings (those that contribute most to the separation of beer and urine samples) and found, among Mass2Motifs more prevalent in beer, choline, aminohexose, and hexose conjugation substructures (also highlighted in Figure 3), and among those more prevalent in urine, creatinine, trimethylamine, and acylcarnitine substructures. The relevance of these metabolite groups to beer and urine highlights the interpretability of the MS2LDA+ analysis; we are able to rapidly observe biochemical differences between the sample groups without relying on the identification of individual molecules. In particular, beer derives partly from plant-based metabolomes rich in glycosylated products. During brewing, sugars are released that can react with other parts of the beer metabolite pool, resulting in many glycosylated products. Creatinine, a byproduct of muscle metabolism cleared from serum by the kidneys, is typically found in urine and not in beer, as expected.

**Differential Prevalence of Mass2Motifs.** To further explore the differences highlighted by MS2LDA+, we performed a differential prevalence analysis for each Mass2Motif between the two sample sets. The 20 Mass2Motifs showing the most consistent differential prevalence (based on a *t* test) for each sample were examined (box plots in Figure 4A,B). Of these, we found many Mass2Motifs to which we could assign a substructure or structural feature. Many Mass2Motifs prevalent in beer were based on sugars (hexoses) and typical plant-based structural motifs like cinnamic acid. Likewise, urine prevalent Mass2Motifs include those structurally characterized as proline betaine and trimethylamine substructures. It is also informative to examine the Mass2Motifs that were unchanged between the two sample groups. These included structural families related to the loss of CHOOH (indicative for a free carboxyl group, present in both amino acids and organic acids) and the nucleotide cytosine. Both urine and beer contain similar numbers of molecules related to those structural families. Indeed, both urine and beer contain amino acids and organic acids which both contribute to the carboxyl loss structural family. Note that within this analysis we are making use of information from spectra that would not be identified by classical existing methods. Alternatively, one could look at differences in prevalence of particular molecular families based on the families of identified metabolites. However, such an analysis risks heavy bias: analysis of spectral database contents reveal substantial overlap to a relatively small set of widely abundant metabolites.[21] Quantifying such bias would be near-

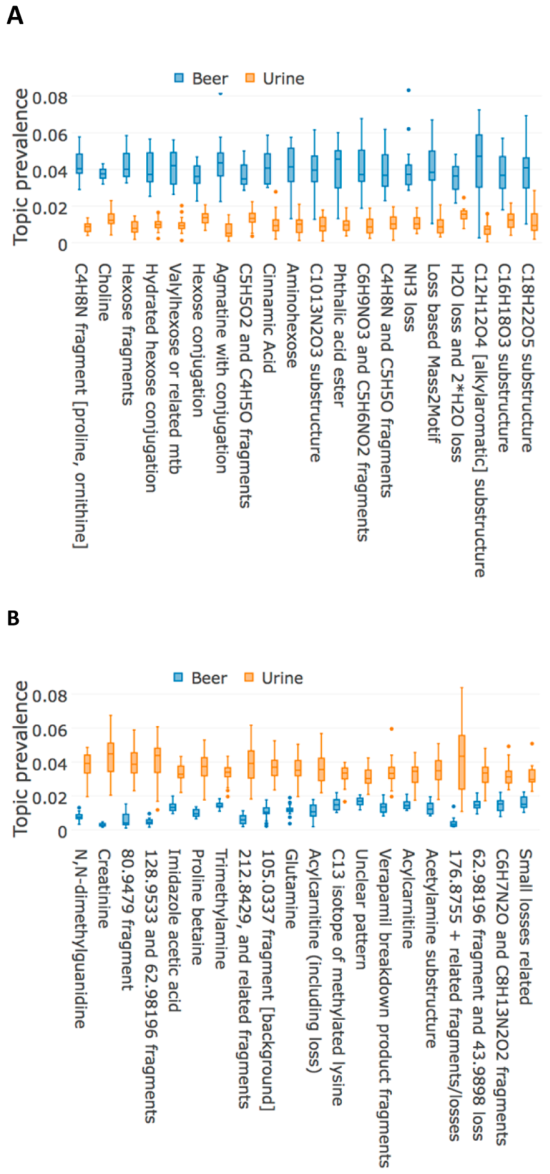


**Figure 4.** Boxplots of high variance alphas abundantly present in beer (A) and urine (B): the top 20 most differential Mass2Motifs are displayed and labeled with their structural characterizations.

impossible. MS2LDA+ sidesteps the problem by considering all spectra that contain a particular Mass2Motif.

These results show that MS2LDA+ can highlight biochemically relevant differences between sample groups in complex metabolomics data. Moreover, we can rank the Mass2Motifs by their variance across the samples and extract the Mass2Motifs that display high variance among the samples. We hypothesize that, within urine samples, those Mass2Motifs that exhibit high variance are likely to represent exogenous metabolites including drugs and food-related metabolites (xenobiotics), while those displaying lower variance are more likely to represent endogenous metabolite families. In the following section we test this hypothesis.

**Mass2Motif Variance within Urines Exposes Xenobiotic-Derived Substructures.** MS2LDA+ was used to discover 300 Mass2Motifs in the 22 urine samples and the $\alpha^f$ values were used to determine the variance for each Mass2Motif (note that to compute the variance we normalized each $\alpha^f$ so that they summed to one over the samples to remove the effect of globally high and low prevalence Mass2Motifs). The 30 most and least variable Mass2Motifs were extracted and, where possible, structurally characterized. Table S6-1 in the Supporting Information, section S6, shows the 30 most variable Mass2Motifs, 23 of which were structurally characterized as containing a biochemically relevant substructure. Nearly all (22) of those substructures can be described as xenobiotic (drug, food-related, or otherwise not naturally occurring in humans). For example among the top 10 most variable Mass2Motifs, we found substructures related to paracetamol mercapturates (a group of metabolites derived from paracetamol, a pain killer (see also Figure 2A–C) for which the Mass2Motif mass fragments matched with the characteristic fragments previously determined from a Molecular Networking cluster,[13] amlodipine (an antihypertensive diuretic), and dimethylated pyrogallol (a marker of polyphenol intake[22]). At the low variance end, 21 out of the 30 least variable Mass2Motifs were structurally characterized as containing a biochemically relevant substructure (Table S6-2 in the Supporting Information, section S6). In the top 10, acetyl loss, cytosine related, and the loss of CHOOH were found, indicating that a constant number of metabolites contain those substructures across the 22 different urines. An additional 57 Mass2Motifs were structurally characterized with biochemically relevant substructures or with related ion products including isotope variants (Table S6-3 in the Supporting Information, section S6). Lysine, carnitine (acylcarnitine), and methyladenine are among those motifs. A total of 101 Mass2Motifs could be structurally characterized, covering on average ∼1600 molecules in a urine sample which represents 86% of the total molecules for which MS/MS data were collected.

Mass2Motifs also show variation across the retention time axis of the mass/charge versus retention time plots (Figure 5). Mass2Motifs like the loss of a pentose (Figure 5B) occur throughout the entire chromatographic window, whereas metabolites containing the sartan-based Mass2Motif elute in a narrow band (Figure 5D). This can further assist in the structural characterization of Mass2Motifs.

## ■ BIOCHEMICALLY RELATED SUBSTRUCTURES CLUSTER BASED ON PREVALENCE ACROSS SAMPLES

(Bio)chemically related substructures (i.e., those created by different pathways derived from the same core metabolite) would likely result in similar $\alpha^f$ profiles across the samples. The most variable Mass2Motif was related to paracetamol mercapturates, containing a variety of mercapturates. Pearson correlation between the $\alpha^f$ vectors for each sample was used to visualize the similarity between the Mass2Motifs profiles with edges between Mass2Motifs where their correlation exceeded a threshold (Supporting Information, section S7). Many xenobiotic Mass2Motifs were found in isolation in the alpha correlation network, indicating their unique prevalence across the urine samples. Conversely, endogenous Mass2Motifs were more densely connected. Interestingly, however, the paracetamol mercapturates Mass2Motif was clustered not only to a Mass2Motif characterized as a paracetamol substructure but also to a Mass2Motif that was subsequently characterized as a group of fragments belonging to methoxyparacetamol. Hence similarities between profiles for individual Mass2Motifs across the samples allow us to prioritize Mass2Motifs to characterize and can assist in the structural elucidation process.

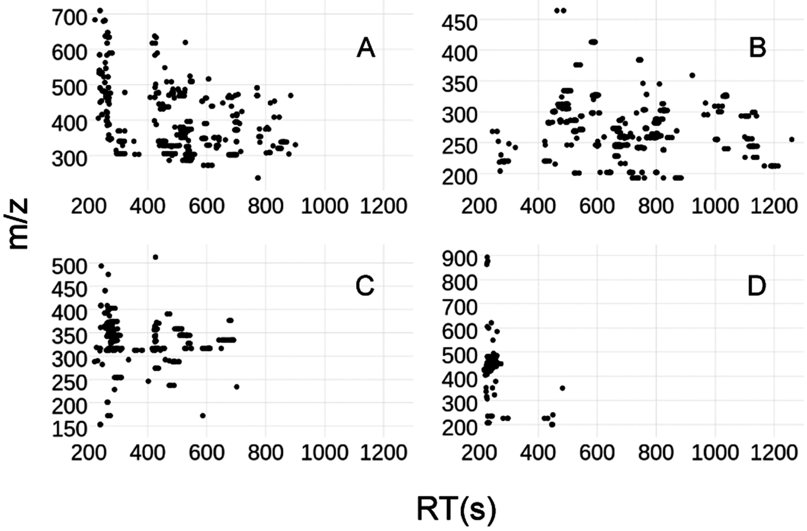


**Figure 5.** Retention time (RT) versus mass/charge ($m/z$) plots for fragmented MS1 ions containing Mass2Motifs annotated as (A) loss of glucuronide, (B) loss of pentose, (C) C10 or longer acylcarnitines, and (D) sartan related drug metabolites. A clear difference in occurrence throughout the chromatogram can be observed with pentose loss occurring throughout the complete RT window while the hydrophobic sartan related drugs are confined to a narrow RT band.

Both paracetamol and methoxyparacetamol were found in urines with two metabolites, namely, the glucuronidated and sulfated form. In our data, these metabolites were characterized by a large number of ion products. In fact, the more paracetamol present in a urine samples, the larger the number of observed ion products. The $\alpha$ values for this Mass2Motif could so be used as a proxy for abundance of those two metabolite families in urine.

In summary, ranking Mass2Motifs by their variance in $\alpha$ guided the extraction of xenobiotic substructures and their associated metabolites in an unsupervised manner. Furthermore, the Mass2Motif $\alpha$ distributions can be used to find biochemically related substructures that can aid in their structural characterization.

## ■ CORE AND RARE METABOLITES GROUPED BY MS2LDA+

A variety of structural families were discovered by MS2LDA+ across the urine profiles in positive mode experiments, including trimethylamine, indole, loss of pentose, loss of glucuronide, glycine, carnitine, and glutamine related Mass2-Motifs. We investigated the carnitine and glutamine families in more details to validate the grouping performed by MS2LDA+. After aligning the MS1 features across the samples, the sample and metabolite normalized MS1 intensities were examined to divide the grouped metabolites into “core” (here defined as present in >80% of urine samples) and “rare” (here defined as present in <80% of urine samples) metabolites. In total, 59 acylcarnitine species were annotated, 47 of which were core acylcarnitines (Supporting Information, section S8). HMDB[23] entries were found for 30% of the annotated core acylcarnitines, whereas none were found for the 12 annotated rare acylcarnitines Supporting Information, section S9). Many of the discovered acylcarnitines were also annotated in previous studies that required more laborious manual annotation.[13,24] This demonstrates a key advantage of working with metabolite families represented by Mass2Motifs, the ability to make use of data that cannot be identified through traditional means.

The acylglutamine family resulted in 23 annotated core acylglutamines (Supporting Information, section S10), of which 3 had hits in HMDB. In total, four nonpeptide glutamine related metabolites are present in HMDB. During annotation of the grouped features, 5 acylglutamine-related ion products were found (isotopes and fragments) and 3 occurrences of cofragmentation of nonacylglutamines. None of the 12 annotated rare acylglutamines (Supporting Information, section S11) had a match in HMDB. An earlier study[13] using Molecular Networking revealed a number of acylglutamine species; however, MS2LDA+ was able to find a larger variety of these species.

Thus, MS2LDA+ allows for mapping of structural families across samples, grouping both features present across all samples as well as those present in just a subset of samples (that would often be discarded in traditional analysis). Performing statistical analysis based on metabolite families allows the use of all related molecules, not just the handful that can be identified by database matching. Only a few acylglutamine species found a match in HMDB, indicating how the grouping by MS2LDA+ aids in structural annotation of features found in untargeted metabolomics experiments. Additionally, the (sample and metabolite) normalized MS1 intensities give insight into abundance and presence/absence patterns within metabolite families which can be grouped within Mass2Motifs using biclusters (Supporting Information, section S12).

## ■ STOOL SAMPLE ANALYSIS OF CHILDREN WITH CROHN’S DISEASE DURING NUTRITIONAL THERAPY

Fecal extracts represent a challenging matrix influenced by many factors such as diet, drug administration, gut microbiota, and the host metabolome. Crohn’s disease is a chronic inflammatory condition of the gut for which no curable treatment is available. To study metabolic differences during disease induction treatment with exclusive enteral nutrition (EEN),[14,15] samples from children with Crohn’s disease and healthy controls were analyzed with MS2LDA+ (see the Supporting Information, section S13). The substructure-based

PCA showed separation of samples taken during treatment versus samples taken after treatment and from healthy controls (Supporting Information, Figure S13A), indicating commonalities between gut substructure contents of healthy controls and patients that had completed treatment. Differential prevalence analysis showed adenine and guanine substructures being depleted in the during-treatment samples, as is also shown in the topic prevalence plots for selected Mass2Motifs (Figure S13B−D).

## DISCUSSION

State-of-the-art mass spectrometers can measure the concentrations and fragmention spectra of small molecules with increasing resolution and coverage, thus improving our chances of discovering biomarkers for the onset of disease and food or drug intake. However, mass spectrometers generate data sets that are very complex and currently tools for analysis use only a small part of the available data.

Unsupervised discovery of substructures in metabolomics data represents a significant step forward in analysis of MS/MS-based metabolomics data.[10] However, armed with that information, the researcher is still tasked with numerous mass fragmental patterns to analyze. Relevant or informative Mass2Motifs are not always present in many metabolites, while small structural families can play crucial roles in biological experiments. Motivated by the observation that Mass2Motifs were similarly defined across different samples, we extended the MS2LDA model to MS2LDA+. MS2LDA+ decomposes multiple samples using a shared set of Mass2Motifs. As well as decomposing each individual molecule into its constituent Mass2Motifs, MS2LDA+ exposes higher level biochemical variability across samples through extracting the sample-specific Mass2Motif prevalence. Differential prevalence and variability of this prevalence can also guide exploration of relevant structural families in metabolomics experiments.

When performing a MS2LDA+ analysis of a collection of beer and urine samples, Mass2Motif prevalence clearly separated the two sample types, demonstrating its ability to extract high-level biochemical changes. Mass2Motifs with high differential prevalence between sample groups made sense in the context of the known biochemical makeup of these complex mixtures. When analyzing the urine samples alone, ranking Mass2Motifs by their variance highlighted xenobiotic Mass2-Motifs whereas low variance Mass2Motifs tended to represent endogenous substructures. For example, the most variable Mass2Motif was related to paracetamol mercapturates, human metabolites of paracetamol (APAP). This highlighted the utility of the MS2LDA+ approach in the extraction of xenobiotic (e.g., drug) substructures (and hence metabolites) from untargeted experiments. Furthermore, MS2LDA+ separated stool samples from children with Crohn's disease during ENN treatment with those after treatment and healthy controls. Substructure analysis revealed significant differential prevalence for a few Mass2Motifs, with adenine and guanine substructures being depleted during treatment. Interestingly, there is evidence that small molecules related to adenine and guanine may play roles in gut microbiota homeostasis and inflammatory response.[25,26] In future, with time-series LC−MS/MS data available for more volunteers, multivariate substructure analysis (explicitly accounting for changes over time) is a promising route to explore. In total, 101 Mass2Motifs were structurally annotated with biochemically relevant substructures. These Mass2Motifs were present in 86% of fragmented molecules (at a cutoff of 0.1 probability score, dropping to 49% at a more stringent cutoff of 0.3), demonstrating the wide coverage that can be obtained in untargeted experiments via the characterization of far fewer Mass2Motifs than original molecules. We are currently considering how to store this structural information in such a way that it can be easily transferred to new experiments and how to make it searchable for the scientific community, for example, by converting them into MassBank records.[27]

The acylcarnitine and acylglutamine metabolite families were investigated in detail and their members partitioned into "core" and "rare" metabolites (Supporting Information sections S8−S11). The grouping of those core metabolites could also allow prioritization of Mass2Motif characterization to those which display large variation in their core metabolite MS1 intensities, as those which are much more likely to produce potential markers than low-variant metabolites or metabolites that are only present in a subset of samples due to limitations of currently used statistical approaches. Biclusters aid in assessment of structural family membership and highlight presence/absence patterns across different samples (Supporting Information, section S12).

## CONCLUSIONS AND OUTLOOK

We introduce MS2LDA+ that provides a platform to guide interpretation of comparative untargeted metabolomics experiments and prioritize structural characterization of Mass2Motifs across large sample sets. We believe that the unsupervised discovery of substructures has particular utility for the detection of *unknown unknowns* (molecules not previously encountered in chemical databases). Moreover, by finding the biochemical relationships between metabolites, we can deal with the gaps in the MS1 data matrix that classical statistical approaches cannot currently do without data imputation strategies.

With beer and urine samples as two distinct groups, we validated that biochemically relevant information is discovered by MS2LDA+, allowing for prioritization of compound classes and groups that contribute to substantial variation in the data. This approach can find relevant substructures and/or substructures of both endogenous and exogeneous origin in urine cohort samples where intragroup variance was suspected to arise from differential drug and/or food administration. In total, we structurally characterized 101 Mass2Motifs discovered in urine. We found that the most variable urinary substructures were xenobiotic-derived.

Substructure-based metabolomics captures most structural information present in fragmentation-enriched untargeted metabolomics experiments. The discovery of structural families in this way will impact many different fields of science. For example, it will be useful in relating lifestyle factors (diet, medicinal use, other exposures) to urinary markers, for monitoring production of natural products in bacteria and plants and monitoring toxic degradation products in wastewater. Moreover, where biochemical pathways are characterized by shared substructures, perturbations in those pathways upon treatments are also likely to be discovered by MS2LDA+ using Mass2Motif differential prevalence analysis. To fulfill the promise of untargeted metabolomics, unsupervised inclusive methods like MS2LDA+ are needed that make use of the biochemical relationships of metabolites and group metabolites per functional group or substructure.

DOI: 10.1021/acs.analchem.7b01391

## ■ ETHICAL APPROVAL

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

## ■ ASSOCIATED CONTENT

### Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b01391.

> Urine samples; beer sample specifications; data acquisition and processing, feature extraction; MS2LDA+ model; overlap score; tables of characterized Mass2Motifs; Pearson correlation between the vectors; table of core acylcarnitine metabolites found across 22 urines; table of rare acylcarnitine metabolites found in subsets of 22 urines; table of core acylglutamine metabolites found across 22 urines; table of rare acylglutamine metabolites found in subsets of 22 urines; normalized MS1 intensities give insight in abundance and presence/absence patterns within metabolite families; stool sample analysis; and references (PDF)
>
> Urine sample data (XLSX)

## ■ AUTHOR INFORMATION

### Corresponding Authors

*E-mail: justin.vanderhooft@glasgow.ac.uk.
*E-mail: simon.rogers@glasgow.ac.uk.

### Notes

The authors declare no competing financial interest.

All data, processed data, and codes used for this paper will be available for download from the university repository (DOI: 10.5525/gla.researchdata.402). In addition, data is available through GNPS/MassIVE: MassIVE data set MSV000081118 contains the urine sample data, MSV000081119 the beer sample data, and MSV000081120 the stool (fecal) sample data. All codes can be found in GitHub (https://github.com/sdrogers/lda).

## ■ ACKNOWLEDGMENTS

J.J.J.v.d.H. was supported by the Wellcome Trust (Grant No. 105614/Z/14/Z). M.P.B. was funded by the Wellcome Trust core grant to the Wellcome Centre for Molecular Parasitology (Grant No. 104111/Z/14/Z). J.W. was supported by a SICSA Ph.D. studentship. K.G. acknowledges support from the Wellcome Trust ISSF funding scheme. S.R. was supported by BBSRC (Grant No. BB/L018616/1).

## ■ REFERENCES

(1) Chaleckis, R.; Murakami, I.; Takada, J.; Kondoh, H.; Yanagida, M. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 4252−4259.

(2) Bouslimani, A.; Melnik, A. V.; Xu, Z.; Amir, A.; da Silva, R. R.; Wang, M.; Bandeira, N.; Alexandrov, T.; Knight, R.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E7645−E7654.

(3) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743−E1752.

(4) Dunn, W. B.; Erban, A.; Weber, R. M.; Creek, D.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. *Metabolomics* **2013**, *9*, 44−66.

(5) da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12549−12550.

(6) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86*, 10724−10731.

(7) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. *J. Nat. Prod.* **2013**, *76*, 1686−1699.

(8) Treutler, H.; Tsugawa, H.; Porzel, A.; Gorzolka, K.; Tissier, A.; Neumann, S.; Balcke, G. U. *Anal. Chem.* **2016**, *88*, 8082−8090.

(9) Misra, B. B.; van der Hooft, J. J. J. *Electrophoresis* **2016**, *37*, 86−110.

(10) van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 13738−13743.

(11) 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.; Crusemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; et al. *Nat. Biotechnol.* **2016**, *34*, 828−837.

(12) Blei, D. M.; Ng, A. Y.; Jordan, M. I. *J. Mach. Learn. Res.* **2003**, *3*, 993−1022.

(13) van der Hooft, J. J. J.; Padmanabhan, S.; Burgess, K. E. V.; Barrett, M. P. *Metabolomics* **2016**, *12*, 1−15.

(14) Quince, C.; Ijaz, U. Z.; Loman, N.; Eren, A. M.; Saulnier, D.; Russell, J.; Haig, S. J.; Calus, S. T.; Quick, J.; Barclay, A.; Bertz, M.; Blaut, M.; Hansen, R.; McGrogan, P.; Russell, R. K.; Edwards, C. A.; Gerasimidis, K. *Am. J. Gastroenterol.* **2015**, *110*, 1718−1729.

(15) Gerasimidis, K.; Bertz, M.; Hanske, L.; Junick, J.; Biskou, O.; Aguilera, M.; Garrick, V.; Russell, R. K.; Blaut, M.; McGrogan, P.; Edwards, C. A. *Inflammatory Bowel Diseases* **2014**, *20*, 861−871.

(16) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. V. *Anal. Chem.* **2011**, *83*, 8703−8710.

(17) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(18) Scheltema, R. A.; Jankevics, A.; Jansen, R. C.; Swertz, M. A.; Breitling, R. *Anal. Chem.* **2011**, *83*, 2786−2793.

(19) Stravs, M. A.; Schymanski, E. L.; Singer, H. P.; Hollender, J. *J. Mass Spectrom.* **2013**, *48*, 89−99.

(20) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(21) Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *TrAC, Trends Anal. Chem.* **2016**, *78*, 23−35.

(22) van der Hooft, J. J. J.; de Vos, R. C. H.; Mihaleva, V.; Bino, R. J.; Ridder, L.; de Roo, N.; Jacobs, D. M.; van Duynhoven, J. P. M.; Vervoort, J. *Anal. Chem.* **2012**, *84*, 7263−7271.

(23) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.

(24) van der Hooft, J. J. J.; Ridder, L.; Barrett, M. P.; Burgess, K. E. V. *Front. Bioeng. Biotechnol.* **2015**, *3*, 3.

(25) Lobo, L. A.; Benjamim, C. F.; Oliveira, A. C. *Clin. Transl. Immunol.* **2016**, *5*, e90.

(26) Idzko, M.; Ferrari, D.; Eltzschig, H. K. *Nature* **2014**, *509*, 310−317.

(27) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; et al. *J. Mass Spectrom.* **2010**, *45*, 703−714.