MetaboShiny - interactive processing, analysis and identification of untargeted metabolomics data

Joanna C. Wolthuis[1,2], Stefania Magnusdottir[1], Mia Pras-Raves[3], Judith J.M. Jans[3], Boudewijn Burgering[1,2], Saskia van Mil[1]*, Jeroen de Ridder[1,2]*

[1]Center for Molecular Medicine, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands

[2]Oncode Institute, Utrecht, The Netherlands

[3]Section Metabolic Diagnostics, Department of Genetics, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands.

*Joint senior authors

Acknowledgements

We thank Arie Kies and Pim Langhout (DSM) for critical discussions, Daphne van Beek and Jasmin Böhmer for assistance with data stewardship, and Marc Pages Gallego for testing and input on the software.

Correspondence should be addressed to:

Jeroen de Ridder, J.deRidder-4@umcutrecht.nl or
Saskia van Mil, S.W.C.vanMil@umcutrecht.nl
UMC Utrecht
Center for Molecular Medicine
STR3.217
PO Box 85060
3508 AB Utrecht
The Netherlands
+31-(0)887550005

Funding contributions

This research is supported by the Dutch Technology Foundation STW, which is the Applied Science Division of NWO, and Technology Programme of the Ministry of Economic Affairs. This research is also supported by DSM Nutritional Products. JdR is supported by a Vidi Fellowship (639.072.715) from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO)

Abstract

Untargeted metabolomics by mass spectrometry in the form of mass over charge and intensity of ions, provides insight into the metabolic activity in a sample and is therefore essential to understand regulation and expression at the protein and transcription level.

Problematically, it is often challenging to analyze untargeted metabolomics data as many m/z values are detected per sample and it is difficult to identify what compound they represent. We aimed to facilitate the process of finding m/z biomarkers through statistical analysis, machine learning and searching for their putative identities.

To address this challenge, we developed MetaboShiny, a novel R and RShiny based metabolomics data analysis package. MetaboShiny features bi/multivariate and temporal statistics, an extensive machine learning module, interactive plotting and result exploration, and compound identification through a variety of chemical databases. As a result, MetaboShiny enables rapid and rigorous analysis of untargeted metabolomics data as well as target identification at unprecedented scale.

To demonstrate its efficacy and ease-of-use, we apply MetaboShiny to a publicly accessible metabolomics dataset generated from the urine of smokers and non-smokers. Replication of the main results of the original publication, which includes importing, normalization and several statistical analyses, is achieved within minutes. Moreover, MetaboShiny enables deeper exploration of the data thereby revealing novel putative biomarkers and hypotheses. For instance, by using MetaboShiny's subsetting feature, iodine is found to be significantly increased in non-smoking lung cancer patients. Furthermore, by allowing for custom adducts, MetaboShiny reveals a putative identification for an m/z value which could not be identified by the original authors. This validates MetaboShiny as a flexible and customizable data analysis package that greatly enhances metabolomics biomarker discovery.

Introduction

Metabolomics is the underlying biochemical layer of the genome, transcriptome and proteome, which reflects all the information expressed and modulated by these omics layers. Because metabolomics provides an almost direct readout of metabolic activity in the organism, metabolomics can be used to diagnose diseases from biofluids, discover new drugs and drug targets, and further precision medicine[1].

A common method to acquire metabolomics data is mass spectrometry (MS), which records the input metabolites' mass to charge ratios (m/z). Targeted MS involves *a priori* metabolite selection based on a chosen compound feature, e.g. polarity, while untargeted metabolomics is performed on the input sample without pre-filtering the sample. An example of an untargeted metabolomics method is direct infusion mass spectrometry (DI-MS), which detects tens to hundreds of thousands of metabolites at single part per million (ppm) accuracy in terms of their m/z values[2]. DI-MS runtimes are in the order of one minute per sample, making it highly suitable for high-throughput applications, such as for instance in diagnostics applications. Problematically, DI-MS routinely produces over a hundred thousand unidentified m/z values, which makes user interpretation of the data exceedingly challenging[2–4].

Before untargeted metabolomics data can be effectively utilized to answer diagnostic or research questions, m/z values need to be matched to a metabolite identity. This is a challenging task and third-party tools that aid in this process are not well developed. A particularly pressing issue is that most current MS analysis tools are limited in the number of databases that are used for metabolite identification, thus limiting the number of metabolite identities that can be matched. For instance, many MS analysis tools rely on a custom built-in database, or a fixed integrated version of a publicly available database, with no opportunity to update to the most recent version[5]. Moreover, identification of the metabolite identity is only one step in the data analysis workflow and needs to be integrated into downstream analytics, such as data normalization, statistical testing, visualization, clustering or supervised machine learning. Many of the currently available tools are only geared towards a subset of these tasks, impeding the ease-of-use and time-to-results.

Given the increasing popularity of untargeted metabolomics we sought to address these issues and enhance the strengths of the available tools and packages. The resulting software package, MetaboShiny, supports the user in performing a range of common MS data analysis steps such as normalization, statistical analysis and putative identification of m/z values. MetaboShiny is built in the interactive *Shiny* framework for the R programming language[6]. MetaboShiny, through our included companion package *MetaDBparse* developed as part of the project, supports 17 databases (Table S1) and is capable of searching through millions of m/z values, including metabolite variants with over twenty commonly observed adducts and isotopes (Table S2, S3). Additionally, users can create their own user-defined databases. Identification and statistical analysis can be performed in parallel, allowing the user to quickly find putative identities for m/z values.

MetaboShiny supports mapping m/z ratios to metabolite identities: matching the highly accurate m/z value to a database (e.g., HMDB, ChEBI, PubChem, etc.) or deducing a molecular formula based on organic chemistry and pre-defined rules such as the *Seven Golden Rules* that take into account ratios of oxygen, nitrogen and hydrogen atoms[3,7].

MetaboShiny performs downstream data analysis directly on the m/z ratios and metabolite identification occurs only on e.g. significant hits that result from the analysis. This has the advantage that the identification remains a separate entity that does not interfere with the original data and statistical analysis, so users remain aware of the untargeted nature of their data. As an alternative, the simplified 'database only' mode allows users to manually fill in a m/z of choice and search through the databases directly.

For statistical analysis, MetaboShiny integrates with the *MetaboAnalystR* package, the backend of the well known MetaboAnalyst, and many other R cheminformatics and statistics packages[8]. Moreover, MetaboShiny offers rapid data subsetting based on user-defined metadata, which can be used to perform analyses on subpopulations of interest, enhancing data reusability and flexible analyses. It also houses the full capabilities of the machine learning R package *caret* to create and analyse predictive models, as the ability of a m/z value to predict a phenotype can be of great interest in biomarker discovery.

Users can seamlessly switch between various visualizations and interactively explore tables and plots, while using various linked databases to find possible identifications for these molecules. As a result, MetaboShiny reduces the need to browse various websites with different interfaces to manually search for m/z value matches. Furthermore, the resulting high-quality and customizable figures are publication ready. Thus, researchers can devote more time to explore putative matches in depth. The installation of MetaboShiny is facilitated by Docker[9], which pre-installs all libraries and packages, and all included databases generated with a single button click and a one-time-only run.

Results

An interactive software suite for untargeted metabolomics analysis

MetaboShiny is an R based application, wrapped in a *Shiny* user interface. It enables users to fluidly switch between running statistical analyses, exploring plots, tables and chemical structures, and m/z identification(*Figure 1*). Furthermore, by summarizing the results in publication ready figures, MetaboShiny decreases the amount of time needed to get from initial data to insights.

The first step of the workflow is normalizing the data. There are various settings for normalization native to MetaboAnalyst, alongside additional options and batch correction[10,11]. MetaboShiny moreover supports many different methods to do statistical analysis. For instance, dimensionality reduction can be achieved through Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA), and users can perform t-tests and fold-change analysis on individual m/z values. Additionally, the user can create volcano plots, heatmaps, and machine learning models supplied by the R package *caret*[12]. Metaboshiny also supports time-series data with options for multivariate and bivariate data(*Figure 1c*).

The user can explore a specific subset of the data, based on the supplied metadata, such as diet, sex, or smoking status(*Figure 1b*). Results from all analyses can be compared through Venn diagrams, which offer hypergeometric testing to test the significance of overlapping hits and are useful for prioritizing multiple significant m/z values. By combining MetaboShiny's data subsetting functionality with Venn diagrams, users can compare the data analyses of different subsets to find specific m/z values of interest.

An important and unique feature of MetaboShiny is that it offers data interactivity through the *DT* and *plotly* R packages[13,14]. The *plotly* and *DT* packages are instrumental in creating plots where users can click and zoom in on specific points or regions of interest or interactive tables that users can filter and sort on, and immediately continue to the m/z value identification step.

Important limitations of currently available m/z identification methods are that only a single or limited set of databases is searched and that searching all available databases is very time consuming. For this reason, MetaboShiny streamlines the process of m/z identification, enabling the user to rapidly retrieve matches in a wide range of compound databases that include adducts and isotopes, according to a predetermined error margin(*Figure 1a, Figure S1)*. Before searching for a molecular identity of an m/z value, users select the databases of interest. MetaboShiny currently supports 16 compound databases, including the Human Metabolome Database (*HMDB*), Kyoto Encyclopedia of Genes and Genomes (*KEGG*), and Chemical Entities of Biological Interest (*ChEBI*)[15–17] (Table S1). MetaboShiny then generates m/z values for common isotopes and adducts known to form in mass spectrometry for each compound in the database (*Figure 1a*). Compound databases are generated and stored locally by the user. MetaboShiny moreover allows users to interactively explore and filter the retrieved matches through interactive summary figures.

MetaboShiny can be downloaded and run either locally or optionally in a multi-user fashion, through the *Docker* platform, facilitating cross-platform use[9]. Source code is available on Github. For ease of use, we recommend Docker to avoid the complications arising from installing dependencies.

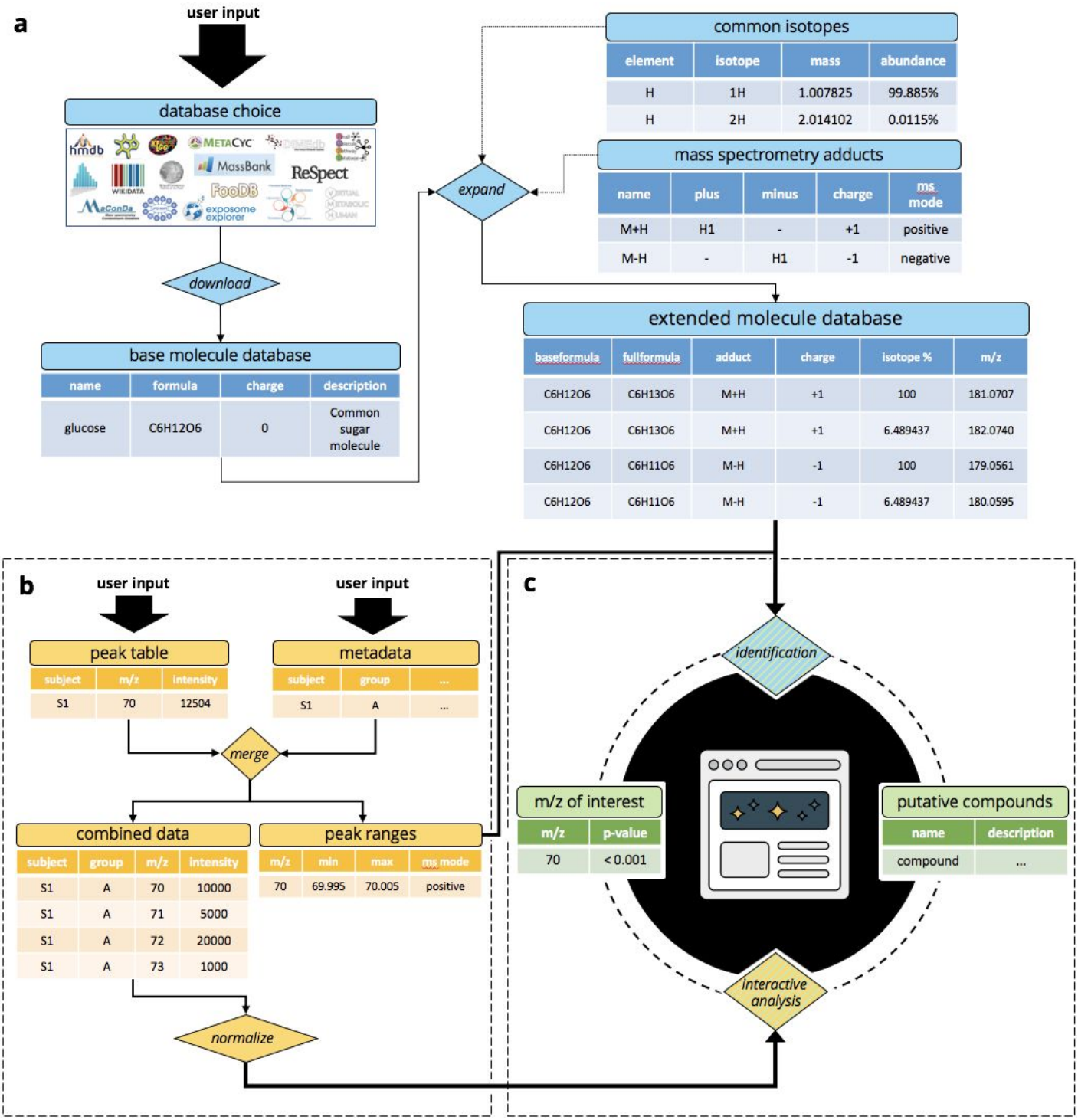


Figure 1: General workflow of MetaboShiny. Users trigger molecule database creation in (a), combine their peak data and metadata in (b) and enter the cycle of biomarker discovery and identification in (c).

Functionality of MetaboShiny demonstrated using a test dataset

To demonstrate the functionality of MetaboShiny, we leveraged the LC-MS dataset on urine samples from 1005 patients with and without lung cancer[18]. This work identified 4 m/z values predictive for cancer status, which were confirmed through targeted mass spectrometry. These data are publically available on MetaboLights[19] (identifier *MTBLS28*).

The available peak tables for positive and negative mode offered a total of 3166 m/z values for each sample. These peak tables were reformatted to remove the retention time, which is currently not searchable in most databases and not used in MetaboShiny for compound identification (script available in supplemental). Additionally, the metadata for each sample was reformatted to fit MetaboShiny's required format. Both files were imported into MetaboShiny. The conversion script and resulting files are available in the electronic supplement. As per the original authors' protocol, an error margin of 25 parts per million (ppm) was used to perform matching[18].

Within MetaboShiny, with a single click after entering normalization preferences, data can be log transformed and normalized based on the sum of the total sample intensity to correct for large differences in overall acquired signal during mass spectrometry detection. We further Z-transformed the data to account for differences between variable distributions. Missing values were imputed with random forest, as it was previously determined to be most accurate[20]. Any m/z values missing a signal in more than one percent of samples were excluded from further analysis. This resulted in 3141 metabolites.

Processing the data from the initial file to the point of statistical analysis takes only 5 minutes once the metadata file is in the right format (which takes approximately 10 minutes, as the MetaboLights metadata format is highly similar to the MetaboShiny format shown in table S4).

Mathé et al. initially trained a random forest model and combined the results of multiple classifiers to find compounds that were good predictors in multiple metadata groups (sex, race, and smoking status)[18]. Using MetaboShiny this analysis can be completely reproduced within only a few minutes, demonstrating the utility of the software for rapid hypothesis generation and biomarker discovery. The Random Forest classifier is amongst the 50+ machine learning algorithms included in Metaboshiny[12]. The model performs better than random classification, with an area under the curve (AUC) of 0.84. Two of the four metabolites that Mathé et al.[18] identified rank within the top 20 in terms of variable importance, with *264.1215224 m/z* having the highest predictive value. Aside from m/z values, smoking status was a strong predictor for disease (ranked #2 by the Random Forest, figure 2j).

When using MetaboShiny to search for potential identifications for the m/z value at the top of the list, a search including the HMDB uncovered creatine riboside (M+H adduct) as a putative hit (*figure 2f).* The built-in HMDB compound description page (*figure 2e)* reports that this m/z feature was first discovered by Mathé *et al*[18] (*figure 2g*), confirming the validity of this finding. All this information can be retrieved with a few mouse clicks.

In addition to the Random Forest model, PLS-DA analysis was explored. This yielded a model that achieved significant separation between the population and lung cancer groups ($p < 0.03$; figure 2h). The loading for PC1 contains the four m/z values of interest, ranked #1, #2, #12 and #15, respectively. These results indicate that the four compounds are not just significant in univariate t-tests, but also can be used to train predictive models that stratify samples by cancer status. These results could be accessed 3 minutes after the PLS-DA analysis was initiated. MetaboShiny can also be used to perform more routine analyses such as t-tests and fold-change analysis.

In depth analysis of test dataset by subsetting and intersecting analyses with MetaboShiny

MetaboShiny enables subsetting and analysis of intersecting features making it very easy for the user to subset and filter samples based on the supplied metadata. For example, as smoking is highly correlated with lung cancer risk, the dataset can be divided into three subsets based on smoking status. For each of these subsets (*present, past* and *never smoker*) regularized (glmnet) predictive models were built with all m/z features (Figure 3a). The predictive importance of each m/z value in these subsets is stored and can be used to do intersectional analysis. The Venn diagram using the top 50 hits for each of these three subsets produced by MetaboShiny is presented in *figure 3* and *table S3*. The analysis revealed that the previously discussed compound creatine riboside (*264.1215224 m/z,* triangle in figures 2 and 3) is in the top 50 predictive m/z values for all three subsets, independent of smoking status *(Figure 3a).* This further validates its status as top significant compound as in the results from Mathe et al[18].

Subsequently, MetaboShiny was used to search for metabolites that only were predictive of cancer status in the group of patients that never smoked. From the top 50 predictive m/z values in total, 44 are uniquely predictive for lung cancer status in this subset, implying a larger role for them in lung cancer not influenced by smoking.

To demonstrate how MetaboShiny enables users to identify putative compounds of interest, these 44 compounds were further explored. As an example, one metabolite of interest, at *126.9069343 m/z,* was identified through browsing the possible m/z identities for metabolites found in humans through the HMDB. MetaboShiny features a side panel showing differential expression for each m/z selected (figure 3b). Selecting this m/z value reveals that the t-test fails to reach significance (indicated by an absence of a star in the figure), indicating that expression in non-smoking lung cancer patients is not significantly higher compared to the healthy population. This highlights the ability of machine learning methods to prioritize m/z values using a different method than conventional statistical tests. Browsing the available identities gave insight into which compounds are interesting for further validation in targeted analyses.

To view the data from another angle, the data was analysed using the machine learning functionality of MetaboShiny. Machine learning models are capable of capturing multivariate variance that other univariate tests cannot. The full *caret* functionality included in MetaboShiny allows users to explore multiple machine learning models, including the random forest, LASSO, ridge regression, general linear models and many more[12].

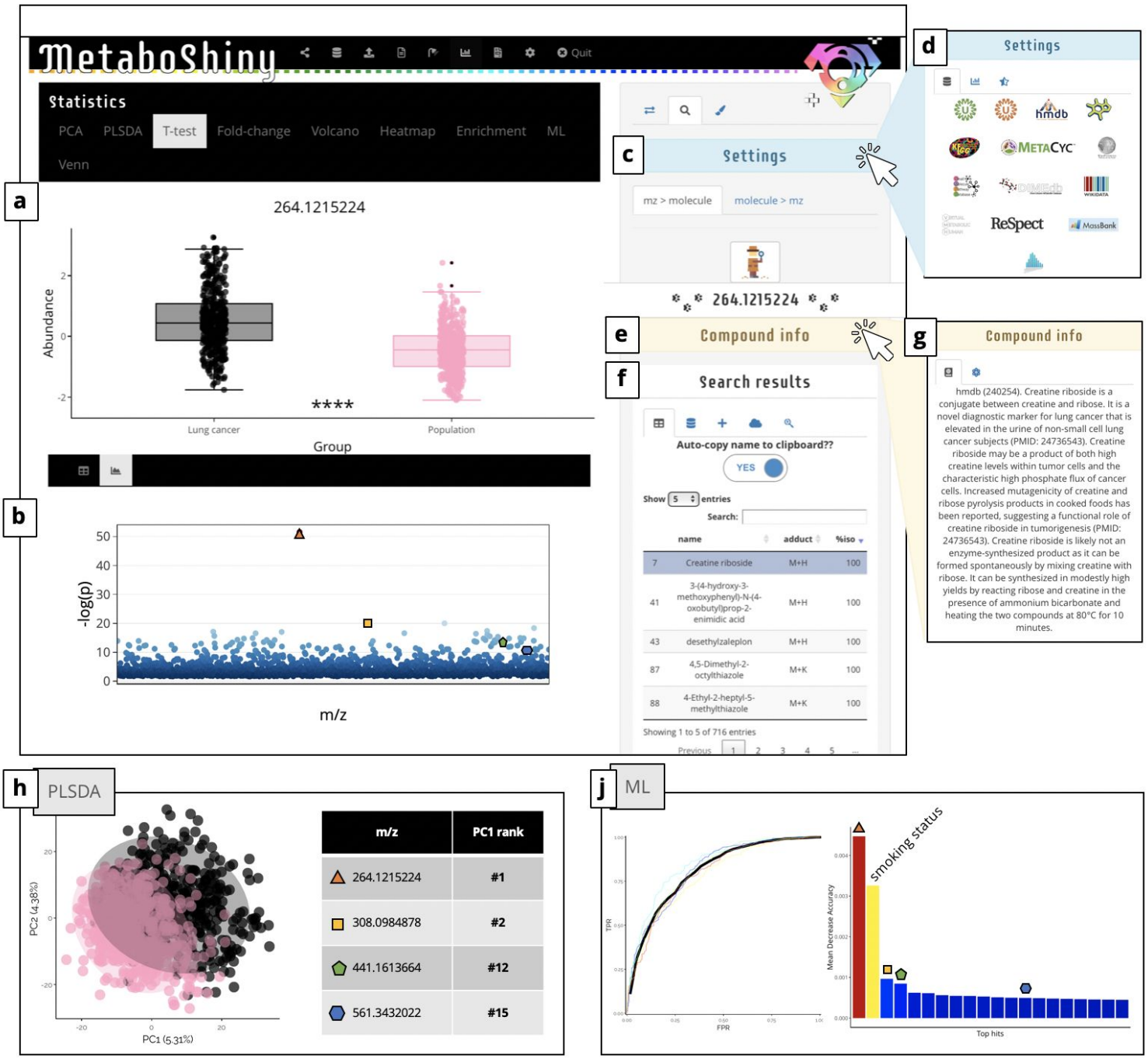


Figure 2: The four previously m/z values found as significant by Mathé et al[18] are highly ranked in multiple statistical analyses within MetaboShiny. (a) Box / beeswarm plot of differential expression of *264.1215224 m/z.* (b) Manhattan plot of all t-test results. X-axis is in order of increasing m/z, y-axis is the -log-transformed p-value. (c-d) Collapsible settings pane including search database selection. (e-g) Collapsible additional information on the currently selected search hit. (f) Table with search hits from the selected databases. (h) PLS-DA results. Table represents the loadings of compounds on PC1. (i) Result of five random forest models

Next, to find possible compound identifications for *126.9069343 m/z*, a full search for all downloaded databases was performed (supplemental table S1). In total this returned 122 matches. After disregarding duplicates, this m/z value matches iodine and iodine-related compounds (Figure 3c). In this case only M-H adducts were considered, as the m/z value was registered in negative mode, and this is the most prevalent ionized form seen in this mode.

To aid the user in interpreting the results, MetaboShiny offers to do a general PubMed search in this case providing a general idea of the co-occurrence of iodine and cancer in literature. A user-specified (500 in this analysis) number of abstracts is parsed and presented in the form of a table with paper titles and either a bar chart or word cloud of the most frequently occurring terms in these abstracts (Figure 3d). In this way, iodine was shown to be often co-mentioned in abstracts with *thyroid*, *radiation* and *uptake,* among others. This, although we lack metadata on how these patients were treated, may have something to do with iodine in cancer treatment. Now that the user has a quick overview of the current state of knowledge on this compound in this context, they can decide either to continue researching further leads within MetaboShiny, or move on to more focused literature research or follow-up experimentation, based on the potentially interesting leads described by the word cloud/ bar chart.

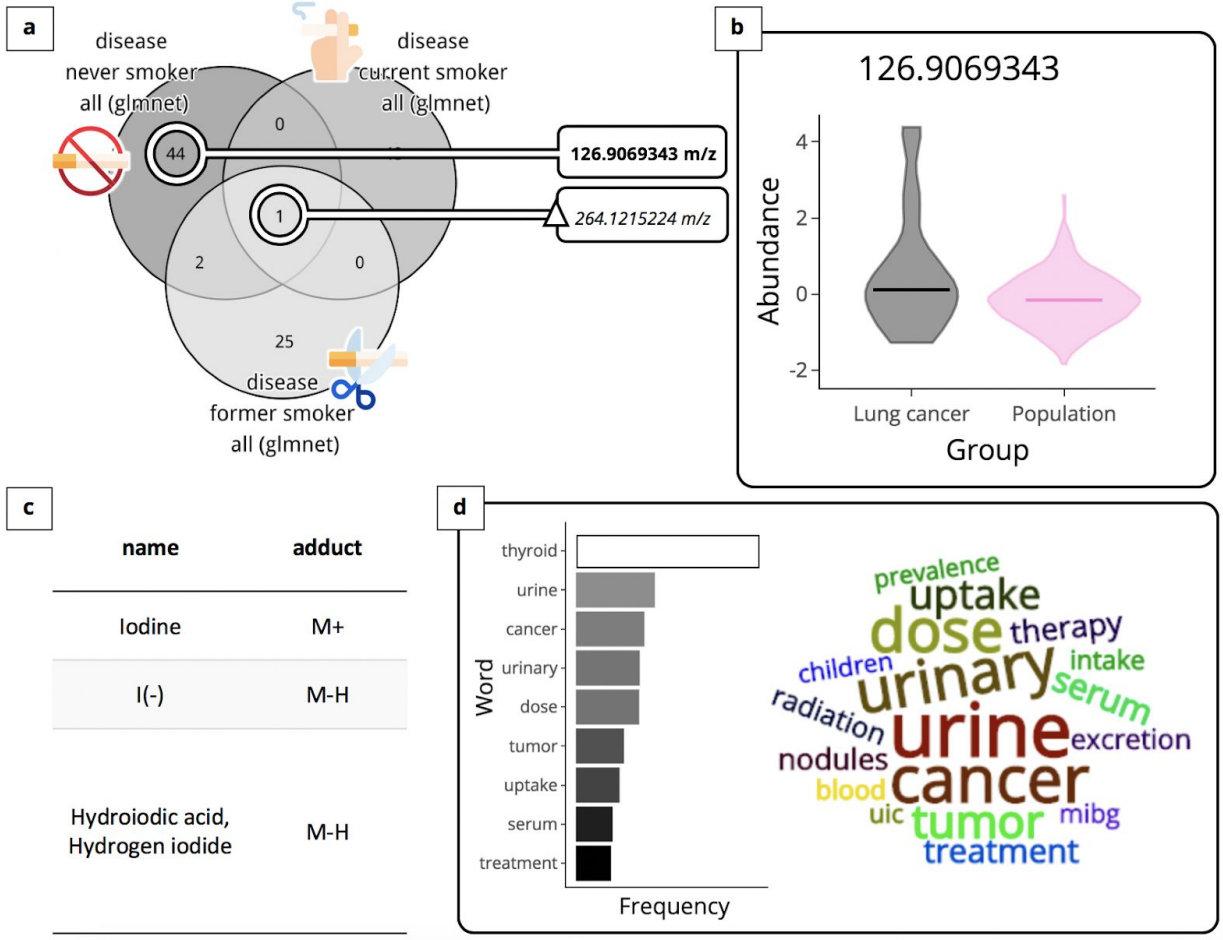
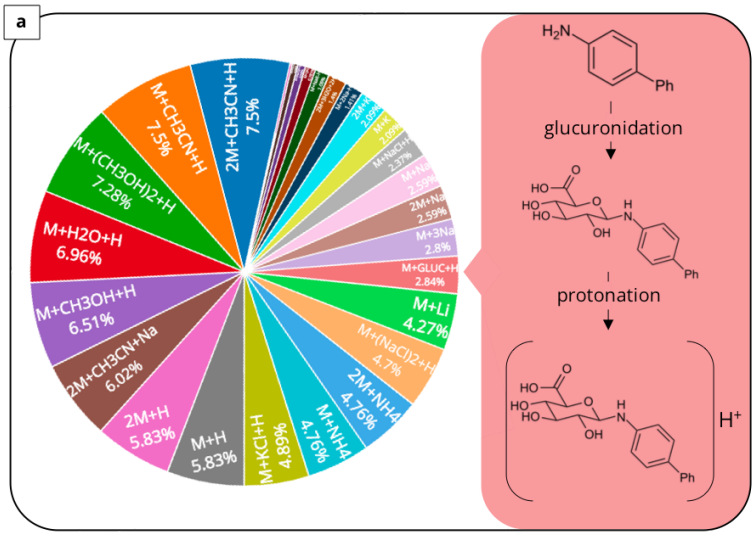


Figure 3: Intersecting the top 50 predictive hits for smokers, nonsmokers and past smokers. (a) Venn diagram of overlap between the top 50 compounds included in final glmnet models for each subset. Metabolites of interest are circled. (b) Differential expression of 126.9069343 m/z between lung cancer patients and population in nonsmokers. (c) Main search results in all databases. (d) Bar chart and word cloud of 500 PubMed abstracts matching the keywords "*iodine cancer urine*".

Matching of a yet unidentified m/z *561.3432022* using Metaboshiny

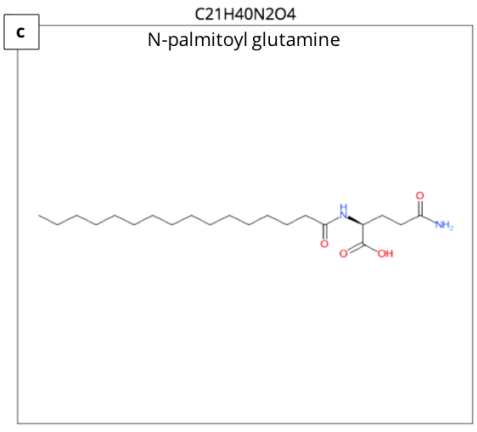
To identify the unknown metabolite with *m/z 561.3432022* the matching capability of MetaboShiny was leveraged. This compound was found to be significantly elevated in lung cancer patients by the original authors[18]. While they established that the compound was conjugated with glucuronide (a process summarized in figure 4a, generally to increase solubility and thus excretability[21]), its exact identity was not unraveled.

A new adduct, named *[M+GLUC±H],* was included in the search to represent glucuronidated compounds. To do this, knowledge on which chemical groups are necessary for glucuronidation was utilized to create rules in SMARTS format(*Table S3*)[21,22]. Following the establishment of this new adduct, MetaboShiny's ability to flexibly calculate molecular adducts was used to implement it in the database(figure 4a). All available databases returned their possible matches for *561.3432022 m/z*. When returning results, MetaboShiny offers the user a pie chart demonstrating from which databases the matches originate, and which adducts were mainly found as possible identities.



b

name	adduct	%iso
N-palmitoyl glutamine	**M+GLUC+H**	**100**
L-759,633	M+GLUC+H	100
L-759,656	M+GLUC+H	100
sesquicannabigerol	M+GLUC+H	100
ethyl lauroyl arginate	M+GLUC+H	100
5-Doxylstearate	M+GLUC+H	100
16-Nitroxystearate	M+GLUC+H	100
BZLDGZBNMOECGO-UHFFFAOYSA-N	M+GLUC+H	100



d

pubmed ID	title
28594046	Balancing the intermolecular forces in peptide amphiphiles for controlling self-assembly transitions.
28413374	**Discrimination of pancreatic cancer and pancreatitis by LC-MS metabolomics.**

Figure 4: Finding an identity for 561.3432022 m/z, a glucuronidated compound. (a) Overview of all hits found for this m/z value shows that 2.9% of all hits are glucuronidated compounds. The right-hand box demonstrates glucuronidation. (b) Top hits of 16-database search for glucuronidated compounds. (c) Chemical structure of N-palmitoyl glutamine generated from SMILES structure description. (d) PubChem papers resulting from the integrated search. Used keyword was 'N-palmitoyl glutamic acid'.

Filtering out everything that was not a glucuronidated adduct through the interactive pie chart in *figure 4a* resulted in multiple main-isotope matches, which can be considered as possible matches for *561.3432022 m/z* (figure 4b). When creating new adducts one must consider if this is chemically valid. This is achieved by using the integrated molecular structure visualization (figure 4c) to estimate if the required groups for glucuronidation are present in the candidate molecule, enabled by the *rcdk* package[23]. Browsing through these results revealed N-palmitoyl glutamine.

The integrated PubChem search (figure 4d) furthermore uncovered a paper demonstrating the potential of a very closely related compound, N-palmitoyl glutamic acid, to differentiate pancreatic cancer from pancreatitis using LC-MS metabolomics[24]. Although this m/z value is found in a different setting, the fact that is is found in the context of LC-MS metabolomics in cancer detection makes this a strong lead to investigate further. Taken together, this demonstrates the power of MetaboShiny to achieve rapid hypothesis generation, which can be followed up with additional experiments.

Comparison of Metaboshiny with existing tools

Many tools perform specific roles in metabolomic analyses, such as pathway analysis or compound identification. MetaboShiny performs multiple analysis steps and integrates many smaller tools into one toolbox. MetaboShiny should therefore be compared to toolboxes with a similar multi-level approach [5]. In this section we discuss four existing platforms and compare them to MetaboShiny.

SECIMtools (SouthEast Center for Integrated Metabolomics) is a suite of metabolomics tools that can be run either as a standalone or on the Galaxy platform for bioinformatics [25]. It specializes in LC-MS data processing and offers multiple statistical analyses (heatmaps, PCA, PLS-DA, clustering, ANOVA), machine learning algorithms (random forest and LASSO), and, accordingly, many visualizations to show the results [26]. MetaboShiny also offers these analyses and further expands upon them, with over seventy machine learning methods available through the caret package. Additionally, although SECIMtools offers more visualizations, MetaboShiny, allows the user to interactively click on the plots to immediately continue analysis and identification, which aids in quick exploration of multiple m/z values as illustrated below with the test dataset. SECIMtools does not offer identification of m/z values, meaning that users have to manually identify their features of interest.

MetaboScape is Bruker's commercial platform for untargeted (LC-MS) metabolomics. It, like MetaboShiny, supports data integration, normalization, statistical analysis and identification [27]. Its identification heavily leans on molecular formula prediction, which is available in MetaboShiny as well. MetaboScape offers many visualizations, but users cannot directly interact with them, but need to pre-label metabolites of interest that are then highlighted in further plots, offering some, but limited, interactivity. MetaboScape does not allow the user to pick databases for identification, nor does it allow users to add their own database. Specialized in LC-MS data, it is not built to work with DI-MS type dataset which contains only m/z values.

Netome, developed by the Broad Institute of MIT and Harvard offers many tools, separately in Python or R, or together in the netome platform. It offers clustering and visualization, alongside access to and tweaking of the raw data[28]. As raw data format and thus necessary processing differs greatly based on the platform used to obtain it, MetaboShiny requires the data to be in peak table format (Table S5). Raw data processing can be performed using specialized packages such as *msconvert* for conversion and the broadly used *xcms* package for peak calling and aligning.

MetaboAnalyst is a webserver with an extensive host of statistical analyses[29]. Users are limited to 5000 metabolites, however, DI-MS usually generates up to hundreds of thousands of m/z

values. MetaboShiny employs the *MetaboAnalystR* package[8] for statistical analyses, while extending its functionality with machine learning and more extensive interaction with result tables and plots.

Additionally, MetaboAnalyst offers limited identification capabilities, most of which are connected to pathway analyses, where m/z notation is replaced by a putative identity. This identity is done by matching the m/z to the HMDB[16]. MetaboShiny gives the user more control over this process, and allows the user to keep track of which m/z connects to which putative identity. An overview of the above comparisons can be seen in *Table 1*.

Table 1: comparison of MetaboShiny with other multifunctional metabolomics tools [26,28–30]. *Checkmark* (✓): full support, *tilde (~)*: partial support and *cross (x)*: no support.

	Metabolite identification	Signature discovery	M/z only statistics	Interactive visualization	Run locally	Open-source
SECIMtools	✗	✓	✗	✗	~	~
MetaboScape	~	✗	✓	✓	✗	✗
NetOme	✗	✗	✗	✗	~	~
MetaboAnalyst	~	✓	✗	✓	~	✓
MetaboShiny	✓	✓	✓	✓	✓	✓

Performance quantification of MetaboShiny

Lastly, to quantify MetaboShiny's speed, a quantitative analysis was performed on MetaboShiny's speed performance. MetaboAnalyst has an upper limit of 5000 compounds for each analysis, in contrast to MetaboShiny, where dataset size is limited only by memory and processor speed of the user's machine. This is the advantage of it being run locally through either Docker or R natively.

We recorded the time of a match for a single m/z value in databases of varying sizes. This is particularly relevant as MetaboShiny allows the user to expand their m/z database by adding more data sources which puts a substantial burden on the matching procedure. For an informative comparison, we created 300 databases spanning the range of real world database sizes. For each of these database sizes, a random pool of 100 m/z values, evenly distributed over the 60 - 600 m/z range were used individually to perform a match search.

Although searching larger databases takes a longer time, *figure 5* reveals a connection between time and the number of matches found for each m/z value. M/z values with more matches in the database will take a bit longer to return, as MetaboShiny fetches additional information (description, molecular formula, name) when a match is found. Regardless of this, one search even at the current maximum database size (if the user downloads and builds all currently included databases), does not exceed 0.8 seconds for a single search, often staying below half a second (*Figure 5b)*.

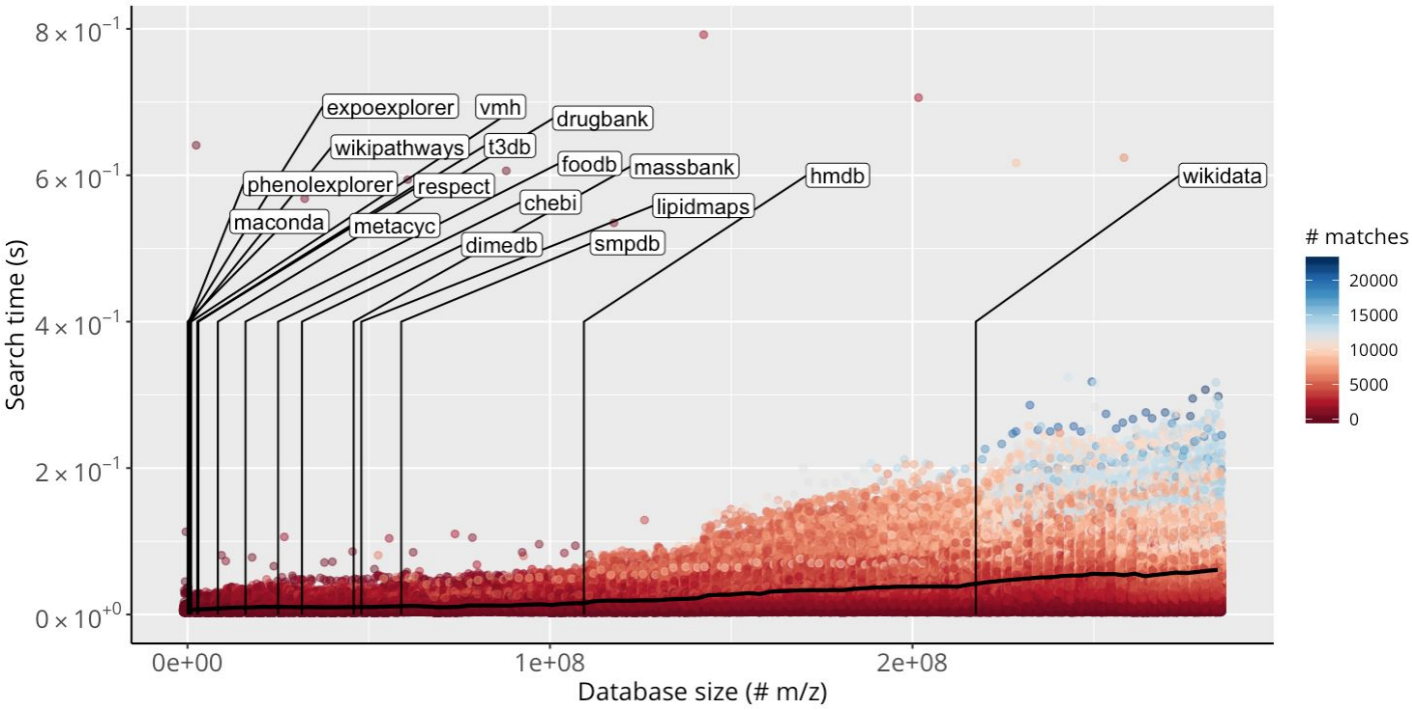


Figure 5: Analysis of MetaboShiny's speed performance. Searching one m/z value takes longer if more matches are found. On average, even with large databases including 3e8 m/z values, performing a search on a single m/z value takes under one second. Labels show the size of various included databases with Wikidata being the most extensive. Color of points represents the amount of matches found for each m/z value.

Discussion

In this paper we introduced MetaboShiny, an R-based and RShiny-based interactive metabolomic analysis and identification tool. It was developed to facilitate metabolomics biomarker discovery, focused on untargeted metabolomics data. MetaboShiny offers a way to integrate large metabolomics datasets with (extensive) metadata, normalize data and perform a broad range of statistical analyses to find potential biomarkers. Once it uncovers m/z values of significance, either through a single analysis or through intersecting the results of multiple analyses, an extensive identification module currently covering 16 databases guides the user towards a putative identity of their compounds of interest, without having to go through multiple databases and websites manually.

We have demonstrated the power of MetaboShiny by reanalyzing the dataset from Mathé et al[18]. Using MetaboShiny, within minutes we were able to validate their number one novel compound of interest, creatine riboside, and show that the other three compounds were all highly ranked in statistical tests. Furthermore, we could perform variations on their Venn diagram analysis combining subset analyses of smokers, non-smokers and former smokers. With integrated subsetting tools, broad-spectrum multi-database- and PubChem searches and various customizable methods of interactive visualization we were able to discover two bona-fide candidates with substantial orthogonal evidence to consider for validation using targeted mass spectrometry. This demonstrates MetaboShiny's ability to guide the user from a peak table and metadata, to biomarkers of interest and their putative identity.

Although we did not find all four compounds at the same significance level as in the primary publication, this is to be expected, since the normalization method and statistical test parameters used here are likely different. We also did not have access to the full metadata including disease stage, which could have been used to reproduce their results more accurately.

Putative identities of m/z values will still require validation through lab work, as untargeted metabolomics only offers m/z values to work with. By prioritizing compounds based on their isotope and adduct status, alongside using database compound descriptions to prioritize compounds based on what is biologically known about them, users can proceed to LC-MS and MS/MS to validate the identity of their marker of interest. Prioritization can potentially be further optimized resulting in fewer potential identities by implementing deep learning using an adapted chemical reaction prediction framework[31].

Like most software tools, MetaboShiny requires data preprocessing. Examples of metadata tables and peak tables compatible with MetaboShiny are given in supplementary table 5 and 6 for metadata and peak data respectively. The metadata format was developed with an in-house data steward to be easy to use. The peak table is required to be in wide format. MetaboShiny does not process raw data output from each type of mass spectrometer and it needs to be processed first.

Further plans include further refining *in silico* compound identification and prioritization and expanding the amount of available databases, further enhancing the ability of MetaboShiny to streamline data integration, statistical analysis and, most of all, compound identity prioritization.

Methods

Databases

The most recent release of the database is either downloaded from the creator or through querying the creator's web application programming interface (API). Each database only needs to be downloaded once but can be updated by re-downloading the latest version of the database. The databases contain the name, molecular formula, charge, description, and if available, structure in SMILES format[32]. Currently, 45 adduct variants, 21 in positive ion mode and 24 in negative ion mode, are calculated for each molecular formula and its isotopes (*Table S2*). Users can add and remove adducts through the settings panel.

In terms of adduct creation, the structure in SMILES can be used to pre-filter which compounds can form which adducts. Users can supply a SMARTS string that finds possible chemical sites needed for the adduct to be created, also called adduct rules[22,32]. If such a rule is not supplied, or no SMILES are available for a compound, the structure is not checked and adduct possibility is based on molecular formula only (presence and absence of necessary atoms for the reaction).

Metadata

When creating a MetaboShiny project, users upload their peak table (in either .CSV or .SQLITE format) and metadata (in .XLS(X) or .CSV format). The general structure of these tables is visualized under the boxes in the directional graph in *Figure 1b*. Examples are provided in *Table S4-S5*. The metadata table requires a date of measurement, sample identifier, unique individual identifier and experimental group. The user can add more data if so desired at will, including other specifics of the experiment such as supplement dosages, diet types or age. All of these features can be used later in subsetting and machine learning. The metadata is included by manipulating the formed *MetaboAnalystR* project (*mSet*) post-normalization within R and can be used to subset data.

Data storage

M/z error margins are stored in SQLITE in an RTree structure, which allows for fast range searches[33]. MetaboShiny uses the mean m/z and parts per million (ppm) error margin to create an m/z range table (min - max m/z allowed to match), which is later used to quickly find matches in the downloaded compound databases. For every sample, the intensity of every m/z peak is stored, alongside the ionization mode it was found in (positive or negative). The ionization mode is important when matching specific adducts, because some adducts are only formed in either negative or positive mode.

The compound databases are also stored in SQLITE. The source database is downloaded from the host and is parsed into a format containing the name, description, formula, charge and structure of each compound using the included *MetaDBparse* companion package. These reformatted 'base' databases are saved separately per source database as separate files. The extended database containing all adducts and isotopes for each molecular structure has a different structure to avoid redundant calculation of adducts and isotopes - it does not use the compound descriptions and names but only hosts the information necessary to calculate m/z values, which is molecular formula, structure and charge. The base and extended databases are joined together when performing identification to provide all the information necessary for the user. The differences in data structure between the base and extended tables are visualised in *Figure 1a*.

Normalization

MetaboAnalystR requires a .CSV file with a specific formatting[8]. For the majority of the normalization options users can refer to the MetaboAnalystR package. MetaboShiny also implements such random forest missing value imputation as additional option[10]. Batch correction is based on the batch included in the metadata file, with or without quality control (QC) samples, using either the *ComBat* or *BatchCorrMetabolomics* packages depending on the presence of these QC samples[11,34].

Visualization

MetaboShiny mostly relies on the *ggplot2* package for data visualization. All of the commonly available color palettes and plot styles from *ggplot2* are available[35]. MetaboShiny features 2d and 3d scatter plots for dimension reduction, line plots (used for temporal analysis and machine learning plots), box/violin/beeswarm plots, heatmaps, pie charts and venn diagrams.

All these plots are interactive because of the *DT* and *plotly* R packages[13,14]. These packages are instrumental in creating plots where users can click on specific points of interest or interactive tables that users can filter and sort on, and immediately continue to the m/z value identification step. This interactivity is particularly useful when exploring heatmaps and volcano plots, where many methods use labels for the hundreds of points or rows. This reduces clutter and users can zoom in on selected areas and explore them at will (Figure S8).

Machine learning

Although MetaboAnalystR natively supports several machine learning methods, MetaboShiny uses the *caret* package[12], which is currently the most extensive machine learning package available for R. It allows users to apply almost a hundred different methods with manually tunable parameters, to maximise flexibility. In terms of results, MetaboShiny offers a summary receiver operating characteristic (ROC) curve, the area under the curve (AUC) and a separate bar chart with variable importance for each m/z and metadata value. These are all interactive, and users can view the weight of each predictor once clicking on a specific ROC curve.

For the analysis discussed in this work, MetaboShiny built 20 random forest models for the analysis of the whole dataset, using parameters *mtry = 2* and *ntree = 500*.

Identification

Every time the user selects an m/z value from a plot or table, it is registered, and this '*current m/z value*' is displayed on the search tab on the sidebar that remains visible during statistical analysis. Once the user has selected adducts and databases, a search can be carried out. Since the m/z value of interest is linked to an m/z range table, which takes the ppm error margin into account, MetaboShiny will retrieve the information associated with all compounds that fall within this m/z range within all selected databases. Results are displayed within a table with multiple pages in the sidebar, alongside a figure summarizing the words used in the descriptions, and two figures summarizing which adducts and which databases are represented in the results. To increase insights, users can choose to search in a user-specified amount of PubMed abstracts and find the most used words in these abstracts. They can also view all of the titles of these papers.

An additional functionality includes predicting possible chemical formulas based on m/z value and predetermined rules, utilizing some the *Seven Golden Rules*[7] (excluding the rules that use presence in existing databases). Using either the elements CHO, CHNO, CHNOP and CHNOPS as possible elements, possible formulas adding up to the m/z in question are generated. These are then filtered according to Golden Rules dictating which ratios of elements to one another are most likely. These rules were acquired from the Visual Basic script offered by the Fiehn lab[7]. Users can choose whether to use these rules or not. Search functionality is provided by our MetaDBparse package which is included in the docker file.

Availability

At time of publication, MetaboShiny will be available on Docker Hub under the *jcwolthuis/metaboshiny* repository. The source code is available on GitHub in the *UMCUGenetics/MetaboShiny* repository. If building from source, administrator rights are needed to install the necessary libraries. The *MetaDBparse* package containing all the search and compound-db specific functionality without the *shiny* user interface is available from the *UMCUGenetics/MetaDBparse* repository.

References

1. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **15**, 473–484 (2016).
2. de Sain-van der Velden, M. G. M. *et al.* Quantification of metabolites in dried blood spots by direct infusion high resolution mass spectrometry. *Anal. Chim. Acta* **979**, 45–50 (2017).
3. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).
4. Lin, L. *et al.* Direct infusion mass spectrometry or liquid chromatography mass spectrometry for human metabonomics? A serum metabonomic study of kidney cancer. *Analyst* **135**, 2970–2978 (2010).
5. Misra, B. B. & Mohapatra, S. Tools and resources for metabolomics research community: A 2017–2018 update. *Electrophoresis* (2018).
6. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. Shiny: web application framework for R. *R package version 0. 11* **1**, 106 (2015).
7. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**, 105 (2007).
8. Chong, J. & Xia, J. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* **34**, 4313–4314 (2018).
9. Anderson, C. Docker [Software engineering]. *IEEE Softw.* **32**, 102–c3 (2015).
10. Stekhoven, D. J. missForest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library* (2015).
11. Wehrens, R. *et al.* Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12**, 88 (2016).
12. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* (2008).
13. Sievert, C., Parmer, C., Hocking, T., Chamberlain, S. & Ram, K. plotly: Create Interactive Web

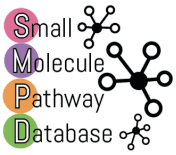
Graphics via 'plotly. js'. *R package version* (2016).

14. Xie, Y. DT: a wrapper of the JavaScript library 'DataTables'. 2016. *R package version 0. 2* (2017).
15. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
16. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
17. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–50 (2008).
18. Mathé, E. A. *et al.* Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res.* **74**, 3259–3270 (2014).
19. Kale, N. S. *et al.* MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr. Protoc. Bioinformatics* **53**, 14–13 (2016).
20. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **8**, 663 (2018).
21. Sanchez, R. I. & Kauffman, F. C. Regulation of Xenobiotic Metabolism in the Liver. *Comprehensive Toxicology* 109–128 (2010). doi:10.1016/b978-0-08-046884-6.01005-8
22. Daylight Chemical Information. SMARTS Theory Manual.
23. Guha, R. & Cherto, M. R. rcdk: Integrating the CDK with R. (2017).
24. Lindahl, A. *et al.* Discrimination of pancreatic cancer and pancreatitis by LC-MS metabolomics. *Metabolomics* **13**, 61 (2017).
25. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* **46**, W537–W544 (2018).
26. Kirpich, A. S. *et al.* SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics* **19**, 151 (2018).
27. MetaboScape 4.0 - Untargeted Metabolomics solution. *Bruker.com* Available at: https://www.bruker.com/products/mass-spectrometry-and-separations/ms-software/metaboscape. (Accessed: 10th May 2019)

28. Rahnavard, A. *et al.* netome: a computational framework for metabolite profiling and omics network analysis. *bioRxiv* 443903 (2018). doi:10.1101/443903
29. Chong, J. *et al.* MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46**, W486–W494 (2018).
30. Edmands, W. M. B., Hayes, J. & Rappaport, S. M. SimExTargId: a comprehensive package for real-time LC-MS data acquisition and analysis. *Bioinformatics* **34**, 3589–3590 (2018).
31. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
32. Anderson, E., Veith, G. D. & Weininger, D. SMILES: A line notation and computerized interpreter for chemical structures. Duluth, MN: US EPA. *Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021* (1987).
33. Allen, G. & Owens, M. SQL for SQLite. in *The Definitive Guide to SQLite* (eds. Allen, G. & Owens, M.) 47–86 (Apress, 2010). doi:10.1007/978-1-4302-3226-1_3
34. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
35. Wickham, H. ggplot2. *WIREs Comp Stat* **3**, 180–185 (2011).

Supplemental Materials

Table S1

Database name	Logo	Description	# unique structures
HMDB		Metabolites commonly found in human biological samples.	11567
ChEBI		A broad database with known chemicals of biological interest.	17727
DimeDB		A direct infusion database of biologically relevant metabolite structures and annotations.	14308
KEGG		Large pathway database with info on pathways in various organisms, involved enzymes, and connected disease phenotypes.	8957
MetaCyc		Large pathway database with over 10.000 available compounds. Spans several organisms.	8386
WikiPathways		Open source biological pathway database. Currently only partially available. Requires CHEBI to be built.	880
SMPDB		Small molecule pathway database. Compounds overlap with HMDB.	2582
Wikidata		Central storage for the data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.	61162

VMH	VIRTUAL METABOLIC HUMAN	Virtual Metabolic Human (VMH) hosts ReconMap, an extensive network of human metabolism, and bacterial metabolites.	3149
ReSpect	ReSpect	RIKEN MSn spectral database for phytochemicals (ReSpect) is a collection of literature and in-house MSn spectra data for research on plant metabolomics.	1604
MassBank	MassBank	This site presents the database of comprehensive, high-resolution mass spectra of metabolites. Supported by the JST-BIRD project, it offers various query methods for standard spectra from Keio Univ., RIKEN PSC, and others.	6822
MetaboLights		MetaboLights is a database for Metabolomics experiments and derived information. The database is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	10059
FooDB	FooDB	FooDB is the world's largest and most comprehensive resource on food constituents, chemistry and biology. It provides information on both macronutrients and micronutrients, including many of the constituents that give foods their flavor, color, taste, texture and aroma.	9092
MaConDa	MaConDa Mass spectrometry Contaminants Database	MaConDa currently contains ca. 200 contaminant records detected across several MS platforms. The majority of records include theoretical as well as experimental MS data.	309
Blood Exposome Database	Precision Medicine, Metabolomics, Exposomics, E, Biomarkers, Risk factors	This new blood exposome database can be applied to prioritize literature-based chemical reviews, developing target assays in exposome research, identifying compounds in untargeted mass spectrometry and biological interpretation in metabolomics data.	23028

LipidMaps		The LIPID MAPS Structure Database (LMSD) is a relational database encompassing structures and annotations of biologically relevant lipids.	7964
Exposome Explorer		Exposome-Explorer is the first database dedicated to biomarkers of exposure to environmental risk factors for diseases.	238
Toxin and Toxin Target Database(T3DB)		The Toxin and Toxin Target Database (T3DB), or, soon to be referred as, the Toxic Exposome Database, is a unique bioinformatics resource that combines detailed toxin data with comprehensive toxin target information.	2658
DrugBank		The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.	10614
Phenol-Explorer		Phenol-Explorer is the first comprehensive database on polyphenol content in foods. The database contains more than 35,000 content values for 500 different polyphenols in over 400 foods.	737
Unique:	**~ 95.000**	**Unique m.z, including. adducts and isotopes**	**> 161 million**

Table S2

Name	Ion_mode	Charge	xM	AddAt	Rem At	AddEx	Rem Ex	Nelec	Rule
[M+3Na]3+	positive	3	1			Na3		-3	Nacc>2 AND Nch=0
[M+H+2Na]3+	positive	3	1			Na2H1		-3	Nacc>2 AND Nch=0
[M+2H+Na]3+	positive	3	1			Na1H2		-3	Nacc>2 AND Nch=0
[M+3H]3+	positive	3	1			H3		-3	Nacc>2 AND Nch=0
[2M+3H2O+2H]2+	positive	2	2			H8O3		-2	Nacc>1 AND Nch=0
[M+3ACN+2H]2+	positive	2	1			C6H11N3		-2	Nacc>1 AND Nch=0
[M+2ACN+2H]2+	positive	2	1			C4H8N2		-2	Nacc>1 AND Nch=0

[M+2Na]2+	positive	2	1			Na2		-2	Nacc>1 AND Nch=0
[M+ACN+2H]2+	positive	2	1			C2H5N1		-2	Nacc>1 AND Nch=0
[M+H+K]2+	positive	2	1			K1H1		-2	Nacc>1 AND Nch=0
[M+H+Na]2+	positive	2	1			Na1H1		-2	Nacc>1 AND Nch=0
[M+H+NH4]2+	positive	2	1			N1H5		-2	Nacc>1 AND Nch=0
[M+2H]2+	positive	2	1			H2		-2	Nacc>1 AND Nch=0
[M+2H-H2O-NH3]2+	positive	2	1		N1H2		O1H1	-2	Nnhh>0 AND Noh>0 AND Nch=0
[2M+ACN+Na]1+	positive	1	2			C2H3Na1N1		-1	Nacc>0 AND Nch=0
[2M+ACN+H]1+	positive	1	2			C2H4N1		-1	Nacc>0 AND Nch=0
[2M+K]1+	positive	1	2			K1		-1	Nacc>0 AND Nch=0
[2M+Na]1+	positive	1	2			Na1		-1	Nacc>0 AND Nch=0
[2M+NH4]1+	positive	1	2			N1H4		-1	Nacc>0 AND Nch=0
[2M+H]1+	positive	1	2			H1		-1	Nacc>0 AND Nch=0
[M+IsoProp+Na+H]1+	positive	1	1			C3H9O1Na1		-1	Nacc>0 AND Nch=0
[M+2ACN+H]1+	positive	1	1			C4H7N2		-1	Nacc>0 AND Nch=0
[M+DMSO+H]1+	positive	1	1			C2H6OS1		-1	Nacc>0 AND Nch=0
[M+2K-H]1+	positive	1	1			K2	H1	-1	Ndon>0 AND Nch=0
[M+ACN+Na]1+	positive	1	1			C2H3Na1N1		-1	Nacc>0 AND Nch=0
[M+IsoProp+H]1+	positive	1	1			C3H9O1Na1		-1	Nacc>0 AND Nch=0
[M+2Na-H]1+	positive	1	1			Na2	H1	-1	Ndon>0 AND Nch=0
[M+ACN+H]1+	positive	1	1			C2H4N1		-1	Nacc>0 AND Nch=0
[M+K]1+	positive	1	1			K1		-1	Nacc>0 AND Nch=0
[M+H+CH3OH]1+	positive	1	1			C1H5O1		-1	Ndon>0 AND Nch=0
[M+Na]1+	positive	1	1			Na1		-1	Nacc>0 AND Nch=0
[M+NH4]1+	positive	1	1			N1H4		-1	Nacc>0 AND Nch=0

[M+H]1+	positive	1	1			H1		-1	Nacc>0 AND Nch=0
[M1+.]1+	positive	1	1					0	Nch=1
[M+H-NH3]1+	positive	1	1		N1H2			-1	Nnhh>0 AND Nch=0
[M+H-H2O]1+	positive	1	1		O1H1			-1	Noh>0 AND Nch=0
[M+H-FA]1+	positive	1	1		C1H1O2			-1	Ncooh>0 AND Nch=0
M	positive	1	1					0	Nch=0
[M+GLUC+H]1+	positive	1	1			C6H10O6	H1	-1	Ngluc>0 AND Nch=0
[3M-H]1-	negative	-1	3				H1	1	Ndon>0 AND Nch=0
[2M+Hac-H]1-	negative	-1	2			C2H3O2		1	Ndon>0 AND Nch=0
[2M+FA-H]1-	negative	-1	2			C1H1O2		1	Ndon>0 AND Nch=0
[2M+Na-2H]1-	negative	-1	2			Na1	H2	1	Ndon>1 AND Nacc>0 AND Nch=0
[2M-H]1-	negative	-1	2				H1	1	Ndon>0 AND Nch=0
[M+TFA-H]1-	negative	-1	1			C2O2F3		1	Ndon>0 AND Nch=0
[M+Br]1-	negative	-1	1			Br1		1	Nacc>0 AND Nch=0
[M+Hac-H]1-	negative	-1	1			C2H3O2		1	Ndon>0 AND Nch=0
[M+FA-H]1-	negative	-1	1			C1H1O2		1	Ndon>0 AND Nch=0
[M+K-2H]1-	negative	-1	1			K1	H2	1	Ndon>1 AND Nacc>0 AND Nch=0
[M+Cl]1-	negative	-1	1			Cl1		1	Nacc>0 AND Nch=0
[M+Na-2H]1-	negative	-1	1			Na1	H2	1	Ndon>1 AND Nacc>0 AND Nch=0
[M1-.]1-	negative	-1	1					0	Nch=-1
[M-H]1-	negative	-1	1				H1	1	Ndon>0 AND Nch=0
[M-2H]2-	negative	-2	1				H2	2	Ndon>1 AND Nch=0
[M-3H]3-	negative	-3	1				H3	3	Ndon>2 AND Nch=0
[M-H+GLUC]1-	negative	-1	1			C6H9O6	H2	1	Ngluc>0 AND Nch=0

Table S3

Short name	Description	SMARTS
Nch	number of charges in M.	*Get charge from SMILES*
Nacc	number of H-bond acceptor in M.	[!$([#6,F,Cl,Br,I,o,s,nX3,#7v5,#15v5,#16v4,#16v6,*+1,*+2,*+3])]
Ndon	number of H-bond donor in M.	[!$([#6,H0,-,-2,-3])]
Noh	number of -OH groups in M.	[OX2H]
Ncooh	number of -COOH groups in M.	[CX3](=O)[OX2H1]
Ncoo	number of -COO- groups in M.	[CX3](=O)[O-]
Nnhh	number of -NH2 groups in M.	[NX3;H2,H1;!$(NC=O)]
Naci	number of acidic H in M.	[H+]
Nbas	number of basic O- in M.	[O-]
Ngluc	number of groups in M that can be glucuronidated	[$([#6][OX2H]),$([OX2H][CX3]=[OX1]),$([NX3,NX4+][CX3](=[OX1])[OX2H,OX1-]),$([NX2:2][OH1:3]),$([NX3]-[CX3](=[OX1])[OH]),$([NX3][CX3](=[OX1])[#6]),$([N;!H0;$(N-c);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]),$([N;!H0;!$(N-c);$(N-C);!$(N-[!#6;!#1]);!$(N-C=[O,N,S])]),$([#16X2H]),$([SX4](=[OX1])(=[OX1])([!O])[NX3]),$([SX4+2]([OX1-])([OX1-])([!O])[NX3]),$([CX3](=O)[CH2][CX3](=O))]

Table S4

Label	Card ID	Animal Internal ID	Sampling Date	Sex	Group
1	CARD1	PAT1	4-Jun-19	Male	Disease
2	CARD2	PAT2	4-Jun-19	Female	Disease
3	CARD3	PAT3	4-Jun-19	Male	Disease
4	CARD4	PAT4	4-Jun-19	Female	Disease
5	CARD5	PAT5	4-Jun-19	Male	Disease
6	CARD6	PAT6	4-Jun-19	Female	Control
7	CARD7	PAT7	4-Jun-19	Male	Control
8	CARD8	PAT8	4-Jun-19	Female	Control
9	CARD9	PAT9	4-Jun-19	Male	Control

Table S5

Sample	71	72	73	74
PAT1	12506	234905	23490	349
PAT2	23940	33494	29385	29485
PAT3	28305	2930	923	20395
PAT4	*NA*	92843	9203	2039
PAT5	3849	*NA*	238940	2189
PAT6	23894	9023	902	23890
PAT7	5863	86739	9283	293
PAT8	82734	29385	2389	67983
PAT9	1283959	38495	1029	23895