Title: Annotating full-scan MS data using tandem MS libraries

Authors:

Shipei Xing[1,2], Vincent Charron-Lamoureux[1,2], Yasin El Abiead[1,2], Pieter C Dorrestein[1,2,3,4,*]

Author affiliations:

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, 92093, USA

[2]Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, 92093, USA

[3]Department of Pharmacology, University of California San Diego, La Jolla, CA, 92093, USA

[4]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, 92093, USA

*Author to whom correspondence should be addressed.

Abstract (70 words, currently 70)

Full-scan mass spectrometry (MS) data from both liquid chromatography (LC) and MS imaging capture multiple ion forms, including their in-source fragments. Here we leverage such fragments to structurally annotate full-scan data from LC-MS or MS imaging by matching against peak intensity scaled tandem MS spectral libraries using precursor-tolerant reverse match scoring. Applied to inflammatory bowel disease and imaging datasets, we show the approach facilitates re-analyses of data in public repositories.

Main (1500 words, currently 2807)

The scientific community is investing substantial resources – funds, time, and effort – in collecting samples, obtaining metabolomics data, analyzing results, and making them publicly available. Just as a conservative number, we are approaching a milestone of one million public liquid chromatography-mass spectrometry (LC-MS) files and over 11,700 MS imaging files across major repositories such as Metabolomics Workbench[1], MetaboLights[2], GNPS/MassIVE[3], and METASPACE[4]. The rationale behind public data deposition (even when restricted access is employed) extends beyond promoting scientific transparency and reproducibility; it also facilitates future reuse, as typically only a small fraction of the data is utilized upon initial publication.

Given that the average annotation rate in metabolomics studies ranges from 10% to 20%[5–8], one crucial aspect of data reuse is to provide new annotations that can be re-contextualized to uncover new biological insights. In untargeted metabolomics, two primary approaches are employed: data collection with tandem MS (MS/MS) and without (MS1-only or full-scan). The latter method offers an advantage in peak shape quality due to the increased number of scans contributing to each peak, leading to enhanced absolute or relative quantification accuracy and thus more reliable statistical analyses. However, MS1-only data presents limitations in discovering molecules that were detected but not yet annotated through subsequent reanalysis. Notably, more than 40% of untargeted LC-MS metabolomics data files in public repositories consist solely of MS1 data, and almost all MS imaging metabolomics data are MS1 scans. This situation creates a significant gap in data reuse and reinvestigation for full-scan data in untargeted metabolomics.

Traditionally, MS1 data interpretation relies on accurate mass measurements and isotopic patterns, which can suggest possible molecular formulas[4,9,10] but often falls short of providing structural information. However, it is also generally understood that many ions may undergo in-source fragmentation or exhibit post-source decay[11–16], generating fragment pieces that also appear in MS1 data. As these processes involve thermal activation, the resulting in-source fragment ions exhibit fragmentation patterns very similar to those observed in collision-induced dissociation (CID) MS/MS spectra. This opens the possibility of leveraging such in/post-source fragments as a handle to create pseudo MS/MS spectra, also referred to as composite spectra[17], that can be leveraged for MS/MS reference library-based annotation in metabolomics and exposomics studies[15–23].

Strategies, such as IDSL.CSA[17], have demonstrated the proof-of-principle of matching pseudo MS/MS spectra, obtained by aggregating ion forms across entire datasets, to reference MS/MS libraries using scoring methods like cosine or entropy similarity[24]. This works particularly well for GC-MS[25] due to the consistent use of a fixed energy (70 eV) for both data acquisition and reference spectra, and the absence of many co-eluting ion forms such as different adducts or

multi-meric species. However, in LC-MS, different ion forms, such as adducts and multimers, often dominate pseudo MS/MS spectra, which may prevent matching to reference MS/MS libraries that do not account for these ion forms. Furthermore, as we show below, the fragment ions we detected tended to match reference MS/MS spectra generated with lower-energy fragmentation, while most reference spectra available in the public domain are collected under medium to high collision energies. In these lower-energy spectra, low-*m/z* ions often appear at very low intensities or may be absent entirely, resulting in missed matches. Other experimental strategies aimed to overcome this limitation–for example, eISA[26] and EISA-EXPOSOME[18] have been developed to incorporate in-source fragments in metabolite annotation. Therefore, in their current implementations, these methods require full-scan data under enhanced in-source CID (isCID) energies (e.g., 40 eV) to obtain adequate spectral matches, and they work in a targeted fashion. Consequently, these methodologies cannot be used for reanalyzing hundreds of thousands of public untargeted MS1 data files acquired without extra isCID experimental designs.

In the realm of MS imaging, metabolite candidate annotations are obtained by annotating molecular formulas and cross-referencing these formulas in common metabolite/lipid chemical databases such as HMDB[27], LipidMaps[28] and ChEBI[29]. This approach corresponds to level 4 identification confidence (unequivocal molecular formula) according to the Metabolomics Standards Initiative[30]. Despite the rapid growth of MS/MS spectral libraries[8], metabolite annotation of MS imaging data has not yet fully benefited from this expanding community resource.

We therefore highlight in this work that we can annotate MS1 data applicable to both LC-MS and MS imaging data. Our approach integrates two steps (although how these steps are implemented is critical): (1) clustering ions or metabolic features through correlation analysis of extracted ion chromatograms (XICs) or ion images in the retention time domain or spatial manner (**Fig. 1a**); and (2) employing a precursor-tolerant (open search[31]) but using reverse spectral matching approach to compare deconvolved MS1 spectra, or pseudo MS/MS spectra, against peak scale-adjusted reference MS/MS libraries for structure candidate identification. Unlike traditional forward spectral matching which utilizes all the peaks in both query and reference spectra for scoring, reverse matching is a unidirectional spectral comparison which ignores unaligned peaks in the query spectrum[32], tolerating contaminant peaks sourced from co-eluting metabolic features or signal artifacts. In the following sections, we elaborate on the spectral matching design and its underlying rationales, demonstrating how this approach enhances annotation capabilities for MS1 data in both LC-MS and MS imaging experiments.

Each molecule detected by mass spectrometry is represented by multiple molecular ions of various adduct forms (e.g., $M+NH_4^+$, $M+Na^+$) that co-elute during chromatography[33–35] or, in the case of MS imaging, share spatial correlations with each other, in addition to in-source fragments. Consequently, intact molecular ions of different adducts, along with their fragments, appear in the

same reconstructed pseudo MS/MS spectrum (**Fig. 1b**). Unlike a typical MS/MS spectrum collected in data-dependent acquisition (DDA) mode, it is not known which ion–if any–represents the precursor ion in a pseudo MS/MS. To address this, we implemented an open search approach–it does not assume a single, predefined precursor ion for each spectrum, but instead considers every ion as a potential precursor ion simultaneously. It employs an unlimited mass tolerance window to accommodate potential mass shifts due to different adducts or multimers, enabling the recognition of various precursor types within the same spectrum.

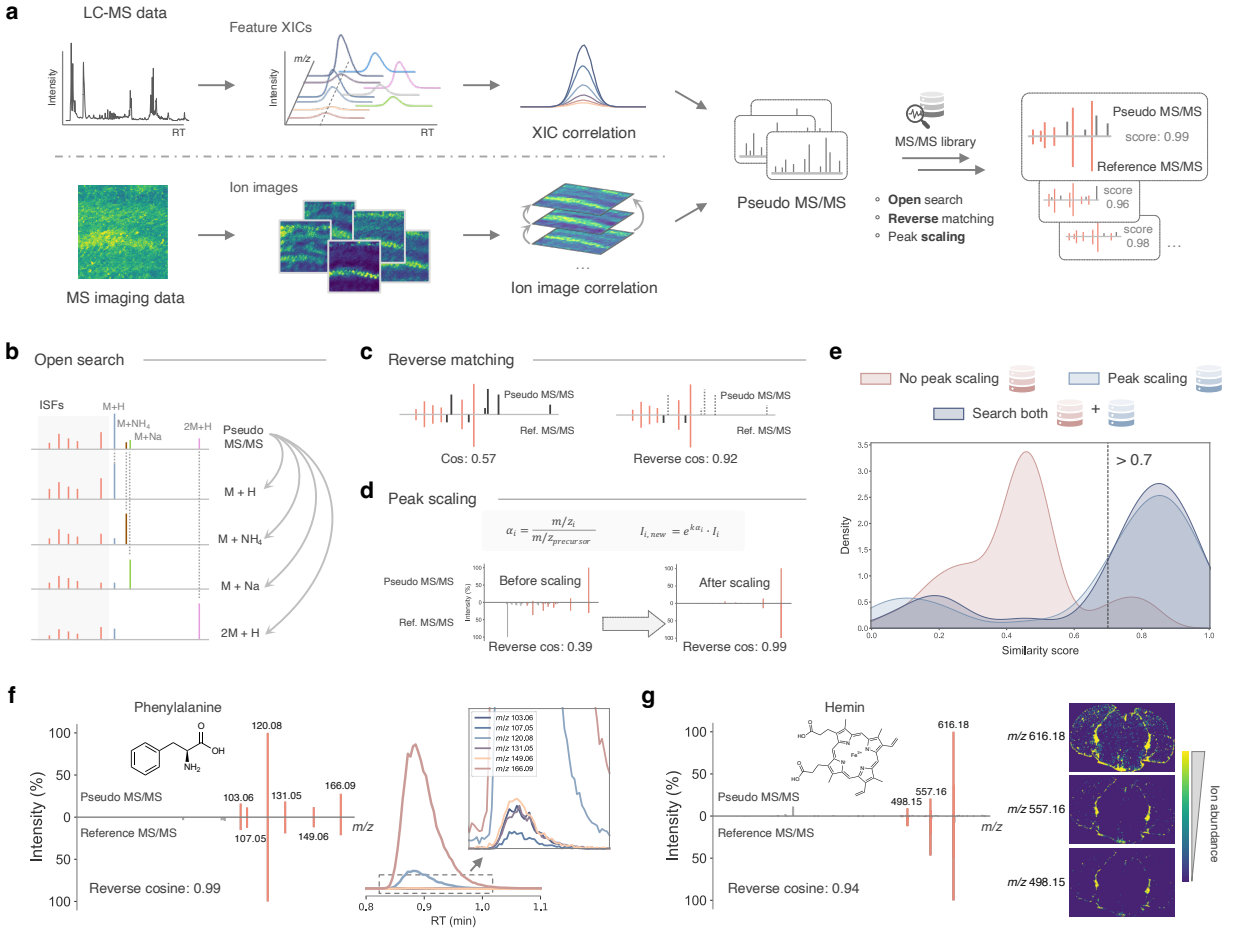


Fig. 1 | Structure annotation of full-scan MS data. **a**, A unified solution to annotate MS1 data from LC-MS or MS imaging experiments. Pseudo MS/MS spectra are generated through correlation analyses in time or spatial domain. The precursor ion-tolerant (open search) reverse spectral searching allows structure annotation of pseudo MS/MS leveraging existing reference MS/MS libraries. **b**, Open search allows matching against reference spectra of multiple adduct forms. ISFs, in-source fragments. **c**, Reverse spectral search discards unmatched peaks in the query spectrum and thus improves spectral search. **d**, Peak intensity scaling helps to match pseudo MS/MS spectra against reference MS/MS which are collected under medium to high collision energies. **e**, Similarity score distribution of searching pseudo MS/MS against libraries with or without peak scaling, or both. Reverse cosine scores of ground truths using chemical standards are used. **f**, An example of structure annotation from LC-MS data (NIST human feces) and XIC correlations. Peak intensities are square rooted. **g**, An example of structure annotation from MS imaging data (mouse brain) and extracted ion images. Ion images were created using 5 ppm mass tolerance. Peak intensities are square rooted.

However, there are challenges in matching a pseudo MS/MS spectrum against the MS/MS spectral library, making it often not possible to provide direct matches. Pseudo MS/MS spectra contain not only fragment ions but also molecular ions of other adduct types and unavoidably mis-clustered ions that are co-eluting. These additional ions are undesirable during spectral matching as they significantly diminish the search scores as the reference libraries do not contain all of the different ion forms. We therefore employed a reverse spectral search (**Fig. 1c**). In this method, reference spectra serve as templates, and unmatched peaks in the pseudo MS/MS are discarded when calculating matching scores. This approach is particularly crucial for MS imaging data, where ions with similar biological functions tend to have similar spatial patterns and thus high correlations, resulting in more ions that should not be compared when trying to annotate (e.g., lipid molecules can exhibit similar spatial distributions on cellular membranes).

Furthermore, as pseudo MS/MS spectra are obtained with minimal energy input (only energy for transfer and/or trapping of the ions), the fragment ion intensities tend to more closely align with low-energy CID spectra. Currently, most reference MS/MS spectra in libraries are collected under medium to high collision energies. Therefore, we developed a peak intensity scaling approach to better align them (**Fig. 1d**). Using chemical standard pools of bile acids and drugs, in total containing 14 known molecules, for which full-scan MS data were collected under in-source CID (isCID) energies of 0 eV, 10 eV and 20 eV, we demonstrated that this peak scaling approach provided more matching scores of >0.7 for ground truths compared to not applying peak scaling (**Fig. 1e**). Combining search results from both original and peak-scaled reference libraries yielded the highest number of matches with reverse cosine scores larger than 0.7.

To further validate our approach, we collected LC-MS data from NIST reference human fecal samples in both DDA and full-scan modes. Full-scan data were acquired under 0 eV, 10 eV, and 20 eV isCID energies. We were able to obtain spectral library matches for 567, 306, 484, 511 and 604 metabolic features in MS/MS (DDA, 42 eV), MS1 (DDA), MS1 (0 eV), MS1 (10 eV), and MS1 (20 eV) modes, respectively (**Fig. 2a**). Unexpectedly, MS1 annotation revealed a unique chemical space, with the majority of annotated features in MS1 data being distinct from MS/MS annotations. More than 79% of the features annotated via pseudo MS/MS lacked corresponding MS/MS spectra in DDA experiments. While DDA typically acquires MS/MS spectra for the more abundant features, this approach captures more low-intensity features when they produce sufficient in-source fragments (**Fig. 2b**). When examining the same metabolic features collected in DDA, structure similarity analyses between MS1 annotations and MS/MS annotations showed that they generated similar chemical candidates (**Fig. 2c**), where a higher isCID energy led to more similar or identical structure matches with MS/MS annotations. We then investigated the compound classes[36] of annotated compounds under different acquisition conditions (**Fig. 2d**). While MS1 data generally annotated more molecules than MS/MS across most compound classes, organic acids & derivatives, and lipids & lipid-like molecules were not as well recognized

in MS1 annotation compared to MS/MS, and this suggests that certain classes of compounds will be easier to annotate via the pseudo MS/MS strategy forwarded here. Overall, above results indicate that MS1 annotation expands the range of detectable metabolites, potentially uncovering previously overlooked compounds in untargeted metabolomics studies, including those available in public repositories.

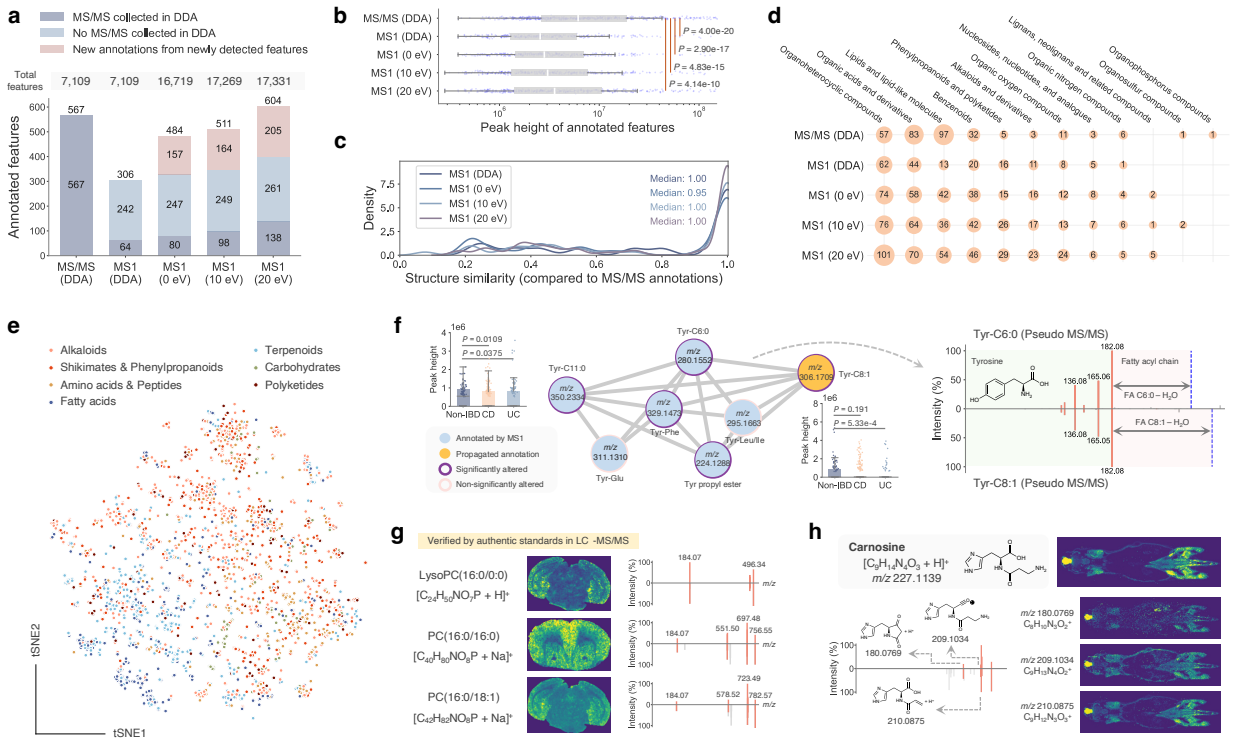


Fig. 2 | MS1 structure annotation provides new insights. **a**, Annotations of NIST fecal samples by MS/MS (DDA) or MS1 data with different isCID energies. **b**, MS1 annotation is able to capture low-abundant metabolic features compared to MS/MS annotations in DDA. Boxes cover the interquantile range (IQR), with medians labeled. The upper whisker represents Q3 + 1.5 × IQR; the lower whisker represents Q1 - 1.5 × IQR. *P* values were calculated using two-sided Mann-Whitney *U* tests. **c**, Structure similarity distributions between MS1 annotations and MS/MS annotations when annotating the same metabolic features. **d**, Compound class distributions of the metabolites annotated in the NIST feces dataset. **e**, t-SNE visualization of MS1 annotations in the IBD dataset. Nodes are colored by compound pathways from NPClassifier. **f**, An example molecular network generated using pseudo MS/MS spectra in the IBD dataset (HILIC positive). Tyrosine-related compounds were annotated and linked. *P* values were calculated using two-sided Mann-Whitney *U* tests. The mirror plot shows pseudo MS/MS spectra from Tyr-C6:0 and Tyr-C8:1. **g**, LysoPC(16:0/0:0), PC(16:0/16:0) and PC(16:0/18:1) annotated in the mouse brain imaging data. They were all verified by authentic standards in LC-MS/MS data. Ion images were created using 5 ppm mass tolerance. Peak intensities are square rooted in the mirror plots, with matched peaks shown in the pseudo MS/MS. **h**, Carnosine annotated in the mouse body imaging data. Ion images were created using 5 ppm mass tolerance. Peak intensities are square rooted in the mirror plot, with matched peaks shown in the pseudo MS/MS.

To highlight reanalysis of a public project with MS information, we revisited a public LC-MS full-scan dataset from an inflammatory bowel disease (IBD) study[37]. This dataset comprised

546 stool samples from three diagnostic groups: non-IBD (n = 134), Crohn's disease (CD, n = 266), and Ulcerative colitis (UC, n = 146). The analysis employed four distinct LC-MS modes: HILIC positive, HILIC negative, C8 positive, and C18 negative. We performed MS1 annotations in batches, obtaining 3010, 293, 227 and 636 unique metabolites (unique InChIKey strings) in the four modes, respectively. Altogether, we identified 3802 unique metabolites with level 2/3 confidence[30]. A t-SNE visualization (**Fig. 2e**), color-coded by compound pathways from NPClassifier[38], revealed distinct clustering patterns among annotated compounds. Alkaloids and shikimates & phenylpropanoids form large, spread-out clusters, suggesting diverse structural variations and their prominence in the gut metabolome. Fatty acids and terpenoids form relatively distinct clusters, indicating unique intensity profiles for lipid-based metabolites across the IBD sample set. We constructed a molecular network using the pseudo MS/MS spectra from the HILIC positive mode. A subnetwork for tyrosine-related compounds was highlighted in **Fig. 2f**, including N-acyl amides that showed alterations in non-IBD vs CD or non-IBD vs UC comparisons. The annotation of Tyr-C8:1 could be propagated through modified cosine-based MS/MS similarity and delta masses with neighbor nodes, and the mirror plot clearly shows the fragmentation pattern of tyrosine as well as the neutral loss of the fatty acyl chain. These findings align with previous IBD studies[39–41], which identified alterations in lipid metabolism and N-acyl amide profiles as key factors in IBD pathogenesis, and highlight that our reanalysis approach can assist in uncovering clinically relevant metabolic signatures the original depositors did not describe.

To demonstrate the efficacy of our strategy on MS imaging data, we applied it onto mouse brain[4], mouse body and human hepatocytes[42] datasets. In the mouse brain sample, we annotated hemin and phosphatidylcholine (PC) lipids of varying chain lengths. **Fig. 1g** displays the ion images of the hemin cation and its in-source fragments, and the visual inspection clearly revealed that the ion image of the hemin cation exhibits spatial patterns highly similar to those of its in-source fragments, with the expected lower abundance of in-source fragments. **Fig. 2g** illustrates the annotations of LysoPC(16:0/0:0), PC(16:0/16:0) and PC(16:0/18:1), which were all verified by authentic standards in LC-MS/MS data[4]. In the mouse body dataset, we obtained 143 candidate annotations. Notably, carnosine was found to be localized to the brain and muscle tissues (**Fig. 2h**), aligning with its dual role as a neuroprotector and a muscle performance enhancer[43]. In the brain, carnosine's presence suggests its involvement in neurotransmitter regulation and synaptic plasticity, processes crucial for learning and memory[44]. In muscle tissue, it functions as an intracellular buffer, regulating pH levels during physical activity, and exhibits antioxidant properties[45] that may aid in recovery from exercise-induced stress. The significant abundance of carnosine in these tissues underscores its importance in both neurocognitive function and physical performance. Extending our analysis to a single cell analysis data set of human cell lines, we examined an MS imaging dataset from differentiated human hepatocytes[42] revealed various lipid classes including phosphatidylcholines, diacylglycerols, and triacylglycerols. As an illustrative example, **Extended Data Fig. 1** showcases the annotation of a diacylglycerol species. This result

highlights the ability of this approach to annotate complex lipids that are interpreted at the single cell level. These findings collectively demonstrate the versatility of our approach across different types of data, from tissue-level imaging to single-cell analysis. By enabling confident annotation of molecular species in various biological contexts, our method promises to enhance our understanding of spatial metabolomics and lipidomics in health and disease.

Despite its capacity to annotate MS1 data from both LC-MS and MS imaging experiments, there are a number of important limitations one has to consider when applying this approach. This approach will not be able to distinguish most isomers, particularly in complex metabolite mixtures with inadequate chromatographic separation. These structurally similar compounds often co-elute and produce similar fragments, impeding the creation of clean pseudo MS/MS spectra and their subsequent distinction. This issue is notably evident in lipid analysis–molecules of the same lipid class share identical characteristic fragments (e.g., the head group ion of phosphatidylcholines), where integration of heuristic rules for retention orders may provide deeper insights. Currently scaling is optimized for maximum number of annotations but this also results in increased incorrect matches compared to no scaling (**Extended Data Fig. 2**). We expect future optimization of scaling can further improve the annotation confidence. Another consideration is the potential for ion contamination or incorrect ion clustering when generating pseudo MS/MS spectra, especially in MS imaging data lacking chromatographic separation. Such limitations elevate the risk of incorrect matches to reference MS/MS spectra. As we show, certain compound classes are underrepresented (e.g., organic acids & derivatives and lipids & lipid-like molecules) in MS1 annotation. This underrepresentation arises from insufficient generation of in-source fragments due to the comparatively low energy imparted on the ions in MS1-only scans. These constraints highlight avenues for future research, including the advancement of more precise MS1 data deconvolution techniques, incorporation of additional orthogonal data for isomer differentiation, and refinement of spectral search algorithms specifically tailored for MS1 data annotation.

Our MS1 annotation approach unveils exciting new prospects for untargeted metabolomics data reuse and analysis. A key opportunity lies in developing an MS1-based MASST[46,47] (Mass Spectrometry Search Tool) to perform reverse metabolomics[48] on LC-MS and MS imaging data, which allows the contextualization of molecules (known or unknown) driven by metadata integration[49] including body distributions, producing organisms, health conditions and interventions. While the current MASST enables searching MS/MS spectra against public data repositories using forward (modified) cosine to retrieve valuable metadata for new biology discovery, MASST could now potentially be extended to the MS1 level. As a proof-of-principle, we queried the pseudo MS/MS spectrum of phenylalanine-C3:0 from the NIST feces sample, which was more abundant in the omnivore group than the vegan group, against the pseudo MS/MS spectra pool from the IBD dataset. This search returned an MS/MS match with cosine score of 0.90 (**Extended Data Fig. 3**). The matched pseudo MS/MS was also annotated as

phenylalanine-C3:0 in the IBD dataset, showing statistical significance in both non-IBD vs CD and non-IBD vs UC comparisons with higher abundance in the non-IBD group. This indicates the feasibility of MS1-based MASST across all four major repositories. Our MS1 annotation approach's ability to identify low-abundance features suggests the possibility of achieving broader metabolome coverage through MS1-based molecular networking[3]. This approach could catalyze the propagation of annotations through spectral similarity analysis, revealing previously unidentified metabolites and facilitating the creation of pseudo MS/MS-based suspect libraries[50] for future data reuse and reanalysis. With over 14,800 untargeted metabolomics datasets (~one million data files) currently available in public repositories, this represents an untapped resource for exploring the dark metabolome[5]–including those elusive metabolites that have thus far escaped identification. As we refine and extend our MS1 annotation techniques, we anticipate an extensive deepening of our understanding of complex metabolic processes and their roles in diverse biological systems and disease states.

Methods

Pseudo MS/MS spectra generation

For LC-MS data, metabolic features are extracted using the MassCube backend[51], which is a Python-based framework for untargeted metabolomics. For each pair of metabolic features within the same retention time window (e.g., ±1.5 s), peak-peak correlation is calculated using their chromatographic profiles. To perform the correlation analysis between two ions, they must share at least 4 consecutive MS1 scans in their chromatographic profiles.

Pseudo MS/MS spectra are then generated as follows: For each metabolic feature (target feature), all other features with correlations exceeding a predefined threshold (e.g., Pearson correlation coefficient ≥ 0.80) are collected. These correlated features are compiled into a pseudo MS/MS spectrum for the target feature. Peak heights of the correlated features in the original MS1 data are used as their respective intensities in the pseudo MS/MS spectrum. Peaks that are determined as isotope peaks by MassCube are excluded from pseudo MS/MS generation.

For MS imaging data, the process is adapted to account for spatial information. Each MS scan undergoes noise reduction using a moving average algorithm. Within a moving window of 100 Da, the baseline is determined as 5 times the mean intensity of the lowest 5% ions in the window, effectively removing background noise. Data centroiding is performed if necessary to reduce data complexity. Ion images are extracted using mass bins of 0.01 *m/z*, and then spatially correlated. A minimum of 5 shared pixels with non-zero intensities between two ion images is required to ensure meaningful correlations and mitigate the impact of sparse data. Pseudo MS/MS spectra are generated by applying a predefined spatial correlation cutoff (e.g., 0.85), followed by deduplication to remove redundant spectra. Both Numba[52] acceleration and parallel processing are employed for computation efficiency enhancement.

Reverse spectral search

Reverse spectral search is an asymmetric matching process, where one spectrum is treated as template (***T***) and the other as query (***Q***). All the peaks in the template spectrum and aligned peaks in the query spectrum are involved in matching score calculation, shown as follows.

$$reverse\ cosine\ = \frac{Q_{aligned} \cdot T_{aligned}}{\left\|Q_{aligned}\right\| \left\|T_{all}\right\|}$$

Considering that pseudo MS/MS spectra are generated from low-energy fragmentation scan modes, and that most public reference MS/MS are acquired under medium to high collision energies, we propose a mass-dependent approach to scale peak intensities for reference MS/MS spectra. This method aims to simulate the pattern observed in low-energy MS/MS, where fragment ions with *m/z* values closer to the precursor *m/z* exhibit higher abundance, while those further from the precursor *m/z* show lower abundance. For a reference MS/MS, we have

$$\alpha_i = \frac{m/z_i}{m/z_{precursor}}$$

$$I_{i,new} = e^{k\alpha_i} \cdot I_i$$

where α_i is the *m/z* ratio of fragment ***i*** to the precursor; I_i is the original intensity; $I_{i,new}$ is the scaled peak intensity; k is the scaling factor where we set it as 8 throughout this paper. Square root transformation is then applied on both reference and pseudo MS/MS spectra. Each pseudo MS/MS is searched against the reference MS/MS library in a precursor-tolerant manner, where we ask that the precursor ions of matching hits should be in the *m/z* values of the query pseudo MS/MS spectrum. For each pseudo MS/MS, we reserve the top 1 hit for each unique precursor *m/z* value among all annotations. Each annotation is then linked to a single metabolic feature in the metabolic feature table using the retention time of the pseudo MS/MS and the precursor mass of the annotation.

To speed up the process of library search, we modified the flash entropy framework[31] specifically for reverse cosine search, which outputs reverse cosine score, matched peak number and spectral usage (sum intensities of matched peaks over total intensities in the query spectrum). The following cutoffs were used for LC-MS MS1 data annotation: minimum score, 0.7; minimum matched peaks, 4; minimum spectral usage, 0.20. For MS imaging data, we used: minimum score, 0.7; minimum matched peaks, 4; minimum spectral usage, 0.05.

The reference MS/MS library needs to be preprocessed and indexed before use. We provided the indexed version of GNPS MS/MS library (downloaded on July 17, 2024) as well as the code to index an MS/MS library on GitHub (https://github.com/Philipbear/ms1_id).

Preparation of chemical standards

For the bile acid pool, a stock solution of 10 mM of glycocholic acid (GCA), glycochenodeoxycholic acid (GCDCA), taurodeoxycholic acid (TDCA), taurocholic acid (TCA), taurolithocholic acid (TLCA), and tauro-3alphahydroxy-12ketocholanoic acid was prepared. All bile acids were diluted to 10 µM into a single 2 mL LC-MS glass vial (Thermo Fisher) to create a pooled sample. For the drug pool, a stock solution of 10 mM of sertraline, venlafaxine, ritonavir, darunavir, losartan, quetiapine, sulfasalazine, and abacavir was prepared. All drugs were diluted to 10 µM into a single 2 mL LC-MS glass vial (Thermo Fisher) to create a pooled sample.

Preparation of NIST reference materials

NIST fecal reference materials (two vegan tubes and two omnivore tubes) were subjected to a biphasic extraction[53] to remove lipids and retain the metabolite fraction. One mL of NIST fecal material was transferred to a 2 mL Eppendorf tube and dried overnight in a CentriVap. Dry materials were resuspended with 325 µL of cold MeOH (LC-MS grade, Thermo Fisher), vortex for 10 s, and sonicated for 5 min before adding 1083 µL of cold MTBE. Samples were vortexed for

10 s and sonicated for 2 min followed by 1 h incubation at 4 °C. To induce phase separation, 271 µL H_2O (LC-MS grade, Thermo Fisher) was added to the samples and centrifuge at 10,000 x g for 10 min. The upper phase was removed and 1084 µL of MeOH was added, followed by an overnight incubation at -20 °C. Samples were centrifuged at 15,000 x g for 10 min. An equal amount (50 µL) of the fecal NIST materials were combined to generate a pooled NIST reference fecal sample. All samples were dried in a CentriVap and stored at -80 °C until resuspension. NIST reference fecal materials were resuspended in 200 µL of 50% MeOH/H_2O with sulfadimethoxine as internal standard before LC-MS analysis.

LC-MS analysis

The chromatographic separation was done on a reverse phase polar C18 (Kinetex Polar C18, 100 mm x 2.1 mm, 2.6 µm, 100 angstrom pore size with the matching guard column, Phenomenex) using a Vanquish UHPLC coupled to an Orbitrap mass spectrometer (Thermo Fisher Scientific). Five microliters of samples were injected into the mobile phase, which is composed of solvent A (H_2O with 0.1% formic acid) and solvent B (ACN with 0.1% formic acid) with the column compartment kept at 40 °C. Samples were eluted at a flow rate of 0.5 mL/min using the following gradient: 0 min, 5% B; 1.1 min, 5% B; 7.5 min, 40% B; 8.5 min, 99% B; 9.5 min, 99% B; 10 min, 5% B; 10.5 min, 5% B; 10.75 min, 99% B; 11.25 min, 99% B; 11.5 min, 5% B; 12.5 min, 5% B. Data were acquired using DDA mode or full-scan mode in electrospray positive ionization mode.

For DDA mode, the parameters were set as: sheath gas flow 53 L/min, aux gas flow rate 14 L/min, sweep gas flow 3 L/min, spray voltage 3.5 kV, inlet capillary 269 °C, aux gas heater 438 °C, S-lens RF level 50.0. MS scan range was set as 100-1000 *m/z* with mass resolution of 35,000 at *m/z* 200. Automatic gain control (AGC) target was set to 1E6 with a maximum injection time of 100 ms. Up to 5 MS/MS spectra per MS1 were collected per cycle with mass resolution 17,500 at *m/z* 200, maximum injection time of 150 ms with an AGC target of 5E5. Isolation window was set to 1 *m/z* and the isolation offset at 0 *m/z*. Stepwised normalized collision energies were set at 25 eV, 40 eV, and 60 eV. The apex trigger was set to 2-15 s and a dynamic exclusion of 5 s. Isotopes were excluded from the analysis.

For full-scan mode, the parameters were set as: sheath gas flow 53 L/min, aux gas flow rate 14 L/min, sweep gas flow 3 L/min, spray voltage 3.5 kV, inlet capillary 269 °C, and aux gas heater 438 °C. MS scan range was set as 100-1000 *m/z* with mass resolution of 70,000 at *m/z* 200. AGC target was set to 1E6 with maximum injection time as 150 ms. Data in full-scan mode were acquired using different isCID energies: 0 eV, 10 eV, and 20 eV.

Statistical analysis

positive mode). For each metabolic feature, we cleaned its pseudo MS/MS by removing all ions larger. For the IBD dataset, missing values were filled using the minimum of 5E5 and 10% of the minimum intensity for each feature. Outlier removal was conducted using the interquartile range (IQR) method. Data points below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR were removed from each group (non-IBD, CD or UC). Two-side Mann-Whitney *U* tests were performed. For t-SNE visualization, intensity values were subjected to log transformation and feature-wise z-normalization.

Molecular networking

A molecular network was constructed using the pseudo MS/MS spectra from the IBD dataset (HILIC positive mode). For each metabolic feature, we cleaned its pseudo MS/MS by removing all ions larger than the feature *m/z*. Then, an MGF file for all cleaned pseudo MS/MS spectra was prepared. The MGF file was uploaded onto the GNPS2 platform, where a classical molecular networking workflow (version 2024.09.20) was completed. A minimum of modified cosine of 0.8 and matched peaks of 6 are required to build an edge in the network construction. The job is available at https://gnps2.org/status?task=670aa34a07544a5cbbd1f1d40605f50f.

Data availability

All the source data used in this study are publicly accessible. For LC-MS data, pooled chemical standards are available at GNPS/MassIVE repository with accession number MSV000095789; NIST human feces are available at GNPS/MassIVE repository with accession number MSV000095787; the IBD dataset is available at Metabolomics Workbench with project number PR000639. For MS imaging data, the mouse brain data are available at MetaboLights repository under code MTBLS313; the mouse body dataset is available at METASPACE platform with ID 2022-07-08_20h45m00s; the hepatocytes data are available at METASPACE platform with project ID Rappez_2021_SpaceM.

Code availability

Source codes are available at GitHub (https://github.com/Philipbear/ms1_id) and Zenodo (https://zenodo.org/records/13864878) under the Apache-2.0 license.

Acknowledgements

This work was supported by BBSRC/NSF award 2152526 and National Institute of Health Sciences U24DK133658. We acknowledge the NIST Complex Microbial Systems Group for providing the NIST material ahead of its official release. Thanks to Theodore Alexandrov for guiding us to the datasets in METASPACE.

Disclosures

P.C.D. is a scientific advisor and holds equity in Sirenas, Cybele, and bileOmix, and is a Scientific Co-founder, and advisor and holds equity in Ometa, Arome, and Enveda with prior approval by UC-San Diego.

Author contributions

P.C.D. and S.X. conceived the research project. S.X. developed the computational algorithm and performed data analysis. V.C.L. collected the LC-MS data. Y.E. provided LC-MS file summaries in public repositories. S.X. and P.C.D. drafted the manuscript. P.C.D. supervised the project. All authors approved the manuscript.

References

1. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
2. Yurekten, O. *et al.* MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* **52**, D640–D646 (2024).
3. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
4. Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods* **14**, 57–60 (2017).
5. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci.* **112**, 12549–12550 (2015).
6. Blaženović, I. *et al.* Structure Annotation of All Mass Spectra in Untargeted Metabolomics. *Anal. Chem.* **91**, 2155–2162 (2019).
7. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 1–20 (2017).
8. Bittremieux, W., Wang, M. & Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* **18**, 94 (2022).
9. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
10. Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat. Methods* 1–10 (2023) doi:10.1038/s41592-023-01850-x.
11. Gabelica, V. & Pauw, E. D. Internal energy and fragmentation of ions produced in

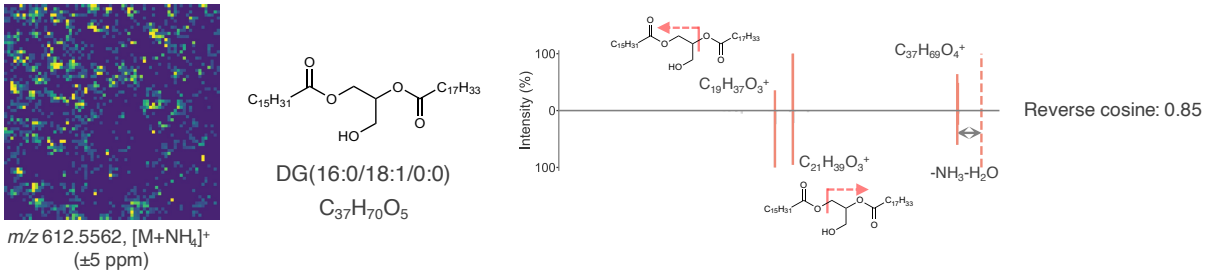
electrospray sources. *Mass Spectrom. Rev.* **24**, 566–587 (2005).
12. Abrankó, L., García-Reyes, J. F. & Molina-Díaz, A. In-source fragmentation and accurate mass analysis of multiclass flavonoid conjugates by electrospray ionization time-of-flight mass spectrometry. *J. Mass Spectrom.* **46**, 478–488 (2011).
13. Criscuolo, A., Zeller, M. & Fedorova, M. Evaluation of Lipid In-Source Fragmentation on Different Orbitrap-based Mass Spectrometers. *J. Am. Soc. Mass Spectrom.* **31**, 463–466 (2020).
14. Xu, Y.-F., Lu, W. & Rabinowitz, J. D. Avoiding Misannotation of In-Source Fragmentation Products as Cellular Metabolites in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **87**, 2273–2281 (2015).
15. Giera, M., Aisporna, A., Uritboonthai, W. & Siuzdak, G. The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation. *Nat. Metab.* 1–2 (2024) doi:10.1038/s42255-024-01076-x.
16. Domingo-Almenara, X. *et al.* Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. *Anal. Chem.* **91**, 3246–3253 (2019).
17. Baygi, S. F., Kumar, Y. & Barupal, D. K. IDSL.CSA: Composite Spectra Analysis for Chemical Annotation of Untargeted Metabolomics Datasets. *Anal. Chem.* **95**, 9480–9487 (2023).
18. Xue, J. *et al.* EISA-EXPOSOME: One Highly Sensitive and Autonomous Exposomic Platform with Enhanced in-Source Fragmentation/Annotation. *Anal. Chem.* **95**, 17228–17237 (2023).
19. Wang, X.-C. *et al.* AntDAS-DDA: A New Platform for Data-Dependent Acquisition Mode-Based Untargeted Metabolomic Profiling Analysis with Advantage of Recognizing Insource Fragment Ions to Improve Compound Identification. *Anal. Chem.* **95**, 638–649 (2023).
20. Wasito, H., Causon, T. & Hann, S. Alternating in-source fragmentation with single-stage high-resolution mass spectrometry with high annotation confidence in non-targeted metabolomics. *Talanta* **236**, 122828 (2022).
21. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).
22. Graça, G. *et al.* Automated Annotation of Untargeted All-Ion Fragmentation LC–MS Metabolomics Data with MetaboAnnotatoR. *Anal. Chem.* **94**, 3446–3455 (2022).
23. Kachman, M. *et al.* Deep annotation of untargeted LC-MS metabolomics data with Binner. *Bioinformatics* **36**, 1801–1806 (2020).
24. Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **18**, 1524–1531 (2021).
25. Aksenov, A. A. *et al.* Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat. Biotechnol.* **39**, 169–173 (2021).
26. Xue, J. *et al.* Enhanced in-Source Fragmentation Annotation Enables Novel Data

Independent Acquisition and Autonomous METLIN Molecular Identification. *Anal. Chem.* **92**, 6051–6059 (2020).

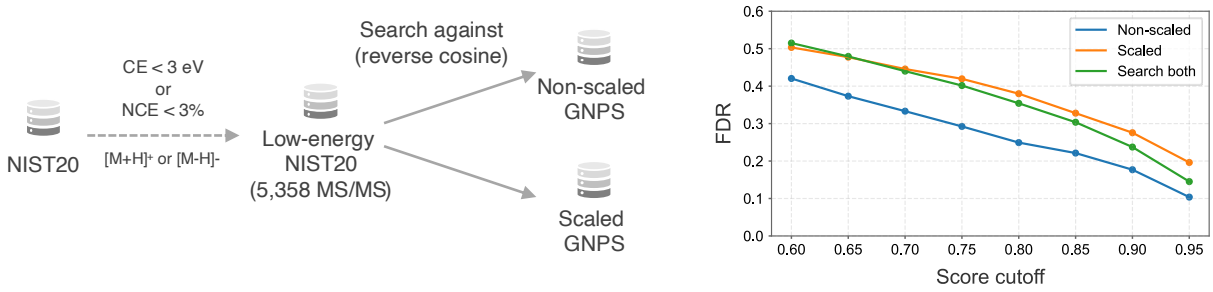
27. Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
28. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**, D527–D532 (2007).
29. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
30. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
31. Li, Y. & Fiehn, O. Flash entropy search to query all mass spectral libraries in real time. *Nat. Methods* 1–4 (2023) doi:10.1038/s41592-023-02012-9.
32. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
33. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
34. Mahieu, N. G. & Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **89**, 10397–10406 (2017).
35. Nash, W. J., Ngere, J. B., Najdekr, L. & Dunn, W. B. Characterization of Electrospray Ionization Complexity in Untargeted Metabolomic Studies. *Anal. Chem.* (2024) doi:10.1021/acs.analchem.4c00966.
36. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, 61 (2016).
37. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
38. Kim, H. W. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **84**, 2795–2807 (2021).
39. Lavelle, A. & Sokol, H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 223–237 (2020).
40. Chang, F.-Y. *et al.* Gut-inhabiting Clostridia build human GPCR ligands by conjugating neurotransmitters with diet- and human-derived fatty acids. *Nat. Microbiol.* **6**, 792–805 (2021).
41. Cohen, L. J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549**, 48–53 (2017).
42. Rappez, L. *et al.* SpaceM reveals metabolic states of single cells. *Nat. Methods* **18**, 799–805 (2021).
43. Derave, W., Everaert, I., Beeckman, S. & Baguet, A. Muscle Carnosine Metabolism and

β-Alanine Supplementation in Relation to Exercise and Training. *Sports Med.* **40**, 247–263 (2010).

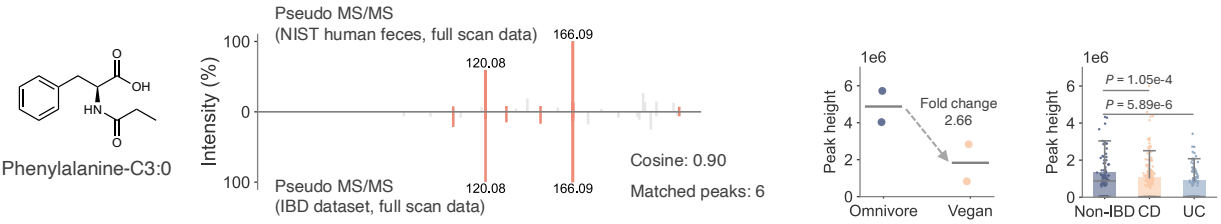
44. Boldyrev, A. A., Aldini, G. & Derave, W. Physiology and Pathophysiology of Carnosine. *Physiol. Rev.* **93**, 1803–1845 (2013).
45. Guney, Y. *et al.* Carnosine may reduce lung injury caused by radiation therapy. *Med. Hypotheses* **66**, 957–959 (2006).
46. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
47. Zuffa, S. *et al.* microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat. Microbiol.* 1–10 (2024) doi:10.1038/s41564-023-01575-9.
48. Gentry, E. C. *et al.* Reverse metabolomics for the discovery of chemical structures from humans. *Nature* 1–8 (2023) doi:10.1038/s41586-023-06906-8.
49. Abiead, Y. E. *et al.* Enabling pan-repository reanalysis for big data science of public metabolomics data. Preprint at https://doi.org/10.26434/chemrxiv-2024-jt46s (2024).
50. Bittremieux, W. *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nat. Commun.* **14**, 8488 (2023).
51. Yu, H. https://github.com/huaxuyu/masscube. (2024).
52. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* 1–6 (Association for Computing Machinery, New York, NY, USA, 2015). doi:10.1145/2833157.2833162.
53. Coman, C. *et al.* Simultaneous Metabolite, Protein, Lipid Extraction (SIMPLEX): A Combinatorial Multimolecular Omics Approach for Systems Biology*. *Mol. Cell. Proteomics* **15**, 1435–1466 (2016).



Extended Data Fig. 1 | DG(16:0/18:1/0:0) was annotated in hepatocytes MS imaging data, with a reverse cosine score of 0.85. Both acyl chains could be annotated in the MS/MS spectrum. The ion image was created using a mass tolerance of 5 ppm.



Extended Data Fig. 2 | Peak scaling tends to increase the false discovery rate (FDR) during spectral matching. We selected low-energy reference MS/MS spectra from the NIST20 MS/MS library using collision energy (CE) < 3 eV or normalized collision energy (NCE) < 3%. These spectra were then searched against the non-scaled and scaled GNPS library. With a minimum of 4 matched peaks (peaks other than the precursor ion), FDR results of different score cutoffs were shown in the line plot. We expect future optimization of scaling can further improve the annotation confidence.



Extended Data Fig. 3 | A proof-of-principle of MS1-based MASST. We searched the pseudo MS/MS spectrum of phenylalanine-C3:0 from NIST human feces data against the pseudo MS/MS spectra pool from the IBD dataset, and it returned a match of cosine 0.90 with 6 matched peaks. The returned hit was also annotated as phenylalanine-C3:0 in the IBD dataset.