**Title:** Charting the Undiscovered Metabolome with Synthetic Multiplexing

**Authors:** Abubaker Patan[1,†], Shipei Xing[1,†], Vincent Charron-Lamoureux[1,†], Zhewen Hu[1], Victoria Deleray[1], Julius Agongo[1], Yasin El Abiead[1], Helena Mannochio-Russo[1], Ipsita Mohanty[1], Harsha Gouda[1], Jasmine Zemlin[1], Prajit Rajkumar[1], Carlynda Lee[1], Daniel Leanos[1], Noah Weimann[1], Wataru Tsuda[1], Sadie Giddings[1], Tammy Bui[1], Kine Eide Kvitne[1], Haoqi Nina Zhao[1], Simone Zuffa[1], Vivian Nguyen[1], Aileen Andrade[1], Wilhan Donizete Gonçalves Nunes[1], Andrés M. Caraballo-Rodríguez[1], Lurian Caetano David[10], Jeremy Carver[9], Nuno Bandeira[1,8,9], Mingxun Wang[7], Lindsey A. Burnett[6], Dionicio Siegel[1,*], Pieter C. Dorrestein[1,3,4,5,*]

**Affiliations**:
[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[2]Department of Medicine, University of California, San Diego, San Diego, CA
[3]Department of Pharmacology, University of California San Diego, La Jolla, CA, 92093, USA
[4]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[5]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, 92093, USA
[6]Department of Obstetrics, Gynecology, and Reproductive Sciences, Division of Urogynecology and Reconstructive Pelvic Surgery, University of California San Diego, La Jolla, CA 92093, USA
[7]Department of Computer Science, University of California Riverside, Riverside, CA, USA
[8]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093, USA
[9]Center for Computational Mass Spectrometry, University of California San Diego
[10]Laboratório de Métodos de Extração e Separação (LAMES), Instituto de Química (IQ), Universidade Federal de Goiás (UFG), Campus II – Samambaia, Goiânia, GO, 74045-155, Brasil

[†] Abubaker Patan, Shipei Xing, and Vincent Charron-Lamoureux contributed equally to this work.

*Co-corresponding authors. drsiegel@health.ucsd.edu for inquiries about synthesis and pdorrestein@health.ucsd.edu for inquiries about metabolomics and analysis.

**Disclosures:** PCD is an advisor and holds equity in Cybele, BileOmix, Sirenas and a scientific co-founder, advisor, holds equity and/or received income from Ometa, Enveda, and Arome with prior approval by UC San Diego. PCD also consulted for DSM animal health in 2023. MW is a co-founder of Ometa Labs.

**Author contributions:** PCD conceptualized the idea. AP led the synthesis part of the project. SX developed the codes for library generation. SX and VCL performed data analysis. AP, SX and JZ created the libraries. YEA, MW developed and managed indexing, metadata harmonization and enabled fast MASST in GNPS2. JC, JD, NB enabled the hardware and software for MassIVE in USIs in GNPS. AP, VCL, JA, and IM performed LC-MS/MS data collection. ZH, VD, DL, TB, SG, NW, VN, WT worked on synthesis. PR developed the SMILES generation tool. CL, AP worked on the SMILES generation. LB obtained urine samples. ZH, VD, AA, and LCD helped with sample preparation. HG, WDGN, HNZ, SZ, KEK, HM-R, AMCR tested synthesis libraries. AP, SX, VCL, and PCD drafted the manuscript. DS and PCD supervised the project. All authors reviewed and approved the manuscript.

## Abstract

In untargeted metabolomics, reference MS/MS libraries are essential for structural annotation, yet currently explain only 6.9% of the more than 1.7 billion MS/MS spectra in public repositories. We hypothesized that many unannotated features arise from simple, biologically plausible transformations of endogenous and exposure-derived compounds. To test this, we created a reference resource by synthesizing over 100,000 compounds using multiplexed reactions that mimic such biochemical transformations. 91% of the compounds synthesized are absent from existing structural databases. Through improvements in the construction of the computational infrastructure that enables pan repository-scale MS/MS comparisons, searching this biologically inspired MS/MS library increased the overall reference-based match rate by 17.4%, yielding over 60 million new matches and raising the global pan-repository MS/MS annotation rate to 8.1%. By facilitating structural hypotheses for previously uncharacterized MS/MS data, this framework expands the accessible detectable biochemical landscape across human, animal, plant, and microbial systems, revealing previously undescribed metabolites such as ibuprofen-carnitine and 5-ASA-phenylpropionic acid conjugates arising from drug–host and host–microbiome co-metabolism.

## Main

Tandem mass spectrometry (MS/MS) spectral reference libraries are essential tools in untargeted metabolomics, enabling researchers to propose plausible structural hypotheses for MS/MS of detected ions of metabolites. Although 93.1% of public MS/MS spectra that are not currently annotated include ion forms such as in-source fragments, different adducts, multimers, or chimerics and low information content spectra (e.g., low signal to background or high intensity but few ions containing spectra)[1–3], the sheer amount of data that remains uncharacterized in publicly deposited metabolomics data likely indicates a significant reservoir of undiscovered biochemistry and uncharacterized metabolic pathways. We hypothesized that many molecules that give rise to these unannotated but detectable ion features result from relatively simple, biologically plausible reactions involving endogenous metabolites and exposure-derived molecules. To enable this expansion of metabolite annotation for which standards are available, we constructed an MS/MS spectral reference library through multiplexed organic synthesis. Reactions were conducted on pools of biologically relevant starting materials, and the resulting mixtures were analyzed via liquid chromatography tandem mass spectrometry (LC-MS/MS) to obtain a synthetic reference library where the MS/MS have known structures. This allowed us to assess whether such molecules had been previously observed in public datasets - a process called *reverse metabolomics* (**Fig. 1a**)[4,5]. Reverse metabolomics involves searching MS/MS spectra across large-scale LC-MS/MS data repositories to identify their occurrence in organisms, organs, health conditions, environments, or other metadata associations available with public data.

This work provides both an annotation library to the community and demonstrates a technological advancement for searching across the ever-growing volume of untargeted metabolomics data. Our earlier work demonstrated the feasibility of reverse metabolomics[4] by synthesizing approximately 2,400 compounds derived from 125 starting materials, including amino acids, bile acids, and lipids[5]. Searching the MS/MS spectra of these molecules across publicly available studies in the GNPS/MassIVE repository[6] at the time resulted in new annotations of approximately 600 molecules. However, the scale of these searches at the time challenged the computational infrastructure, with some searches taking several days/weeks to complete. To enable searches at the scale of hundreds of thousands of MS/MS spectra - generated via multiplexed synthesis - we engineered a system capable of processing MS/MS search against >1.7 billion public MS/MS in milliseconds per query and thousands of queries per minute. This capability was enabled through a combination of hardware upgrades, algorithmic indexing strategies, and software engineering optimization.

The large-scale MS/MS spectral comparisons required for this project required a dedicated expansion and engineering of the computational infrastructure capable of doing so. Reverse metabolomics analyses are now performed on a virtual machine equipped with two 64-core AMD processors and 2 TB of RAM, with public metabolomics data indices hosted on four SSDs to ensure rapid access. This setup supports high-speed spectral searches using indexed spectra - enabling the fast MASST (FASST) queries[7,8]. The second generation GNPS platform[6], the data and knowledge ecosystem that is being searched, has, as this was needed for this project, since expanded to operate across five interconnected virtual machine servers: two equipped with dual 64-core AMD processors with 2 TB of RAM each, and three with dual 16-core CPUs totaling 768 GB of RAM. Data storage is distributed across two high-performance arrays, comprising 424 TB of SSDs - all linked through a 10 Gbit network backbone.

Together, this infrastructure underpins the GNPS2/MASST ecosystem, enabling community-scale reverse metabolomics and repository-wide MS/MS spectral searches at the necessary speed and depth.To broaden the search space, we indexed all data in GNPS/MassIVE[6] and integrated additional public repositories including MetaboLights[9], the Metabolomics Workbench[10] and, more recently, NORMAN, a more environmentally focused repository, via the Pan-ReDU framework[11]. Searches can be performed using fast MASST[7], along with its domain-specific variants (e.g., for microbes[12], food[13], plants[14], tissues[15]). In parallel, we enhanced the underlying data science by continuing to harmonize metadata vocabularies across these repositories, enabling MASST searches to return MS/MS spectral matches and additional relevant and interpretable metadata about the matched samples[11]. All indexed LC-MS/MS files, features, spectra, and synthetic reference libraries were converted to use Universal Spectrum Identifiers (USIs)[16], ensuring complete provenance and traceability to the original deposited raw data. As of late Aug/ early Sept 2025, the number of LC-MS/MS files that are indexed and have harmonized metadata in PanReDU has now grown to 920,790 LC-MS/MS files. This indexed infrastructure supports searches across 4,990 datasets from the repositories previously mentioned, comprising a total of 1,752,167,824 MS/MS spectra. These advancements eliminate previous computational bottlenecks and enhance the biological and environmental interpretability of reverse metabolomics results at scale.

Using this infrastructure, we expanded the chemical space in this study by incorporating a more structurally diverse set of compounds representing a range of biologically and exposure-relevant starting materials. These included 1,450 small molecules that possessed functionality that were plausibly available for biotransformation. The precursor molecules span core metabolic pathways (e.g., central carbon and fatty acid metabolism), steroidal scaffolds (e.g., bile acids), neurotransmitters, vitamins, as well as exposure-related compounds such as dietary components, plastic-associated chemicals, ingredients from personal care products, chemicals used in manufacturing and current approved drugs (**Fig. 1b, Supplementary Table 1**). We prioritized compounds containing amines, carboxylic acids, and hydroxyl groups because their chemical reactivity in biological systems is well-characterized. These functional groups readily form esters, amides, and other common products, which can be readily generated in standard flask-based reactions. This increases the likelihood of identifying relevant reaction products. To model biologically and environmentally relevant biochemical transformations, we applied both single- and multi-step reactions with a multiplexed synthetic strategy, where multiple products are generated with multiple reagents in one reaction vessel (**Fig. 1a**;  details for each reaction can be found in **Supplementary Table 2**). These flask-based reactions were designed to emulate transformations commonly occurring *in vivo* or in the environment, including sulfation, conjugation, methylation, oxidation, hydrolysis, and amide formation (**Fig. 1c, d**). As the objective was to generate detectable products for MS/MS acquisition rather than maximize chemical yield, limited reaction optimization was carried out and no purification steps were needed given we are using chromatographic separation and that due to the sensitivity of mass spectrometry, it is possible to detect and obtain MS/MS data for products that see only a small amount of conversion in the multiplexed reactions. The mixtures with known starting materials, reagents and expected products were analyzed by LC-MS/MS using data-dependent acquisition to capture fragmentation spectra. All resulting mass spectrometry data have been made publicly accessible in the GNPS/MassIVE repository. A total of 492,376 MS/MS spectra were linked to structures, suitable for downstream computational analysis and public data searches via the MASST platform. The remaining MS/MS that are not linked to structures are redundant MS/MS spectra of the same ion forms of molecules, other ions forms that we did not look for (e.g. in-source fragments, different adducts and multimers)[2], chimeric spectra, impurities or unanticipated reactions.

To systematically predict and annotate possible products from the LC-MS/MS data obtained from the multiplexed reactions, we developed a web application for *in silico* generation of all plausible structures of the products from our multiplexed and combinatorial reactions under defined conditions, as existing tools could not sufficiently scale. We then linked the MS/MS to ionic forms of the structures (with $H^+$, $Na^+$, $NH_4^+$, $K^+$ adducts) that could be present in each individual synthetic reaction, including starting materials. This resulted in the 492,376 MS/MS spectra with molecular structures annotations that were synthesized. The outcome is an openly and freely accessible MS/MS reference library. Due to redundant spectra for the same compounds, as well as isomeric overlap (as further discussed in the limitations section), the complete MS/MS  library generated using multiplexing contains 172,483 candidate compounds. Due to structural isomers in the multiplexed reactions, these structural isomers generated in the multiplexed reactions are represented by 134,453 unique MS/MS spectra, each indexed with a USI[16]. This means that when one obtains a match to that particular MS/MS one has to consider all isomers, even ones that might not be present in the synthetic reactions, similar to other MS/MS reference libraries based annotations (**see limitations discussion of this paper**). The candidate molecules that make up the multiplexed library cover diverse chemical classes relevant to both biological and environmental systems (**Fig. 1d**), spanning a mass range from approximately 150 Da to 1,350 Da (**Fig. 1e**). Given the synthesis prioritization of biologically relevant precursor molecules, we anticipated that a significant portion of this library would represent previously unexplored chemical space in biology. Based on the planar structures, 91% of these compounds were unique to our library and not present in any major structure databases (**Fig. 1f**). The highest overlap was with PubChem[17] (8%), which contains over 110 million structures. All other databases, including HMDB[18], GNPS[6], PubChemLite[19], NORMAN and FooDB[20], shared less than 1% overlap with the structures from the multiplexed synthetic MS/MS library (**Fig. 1f, Supplementary Table 3**).

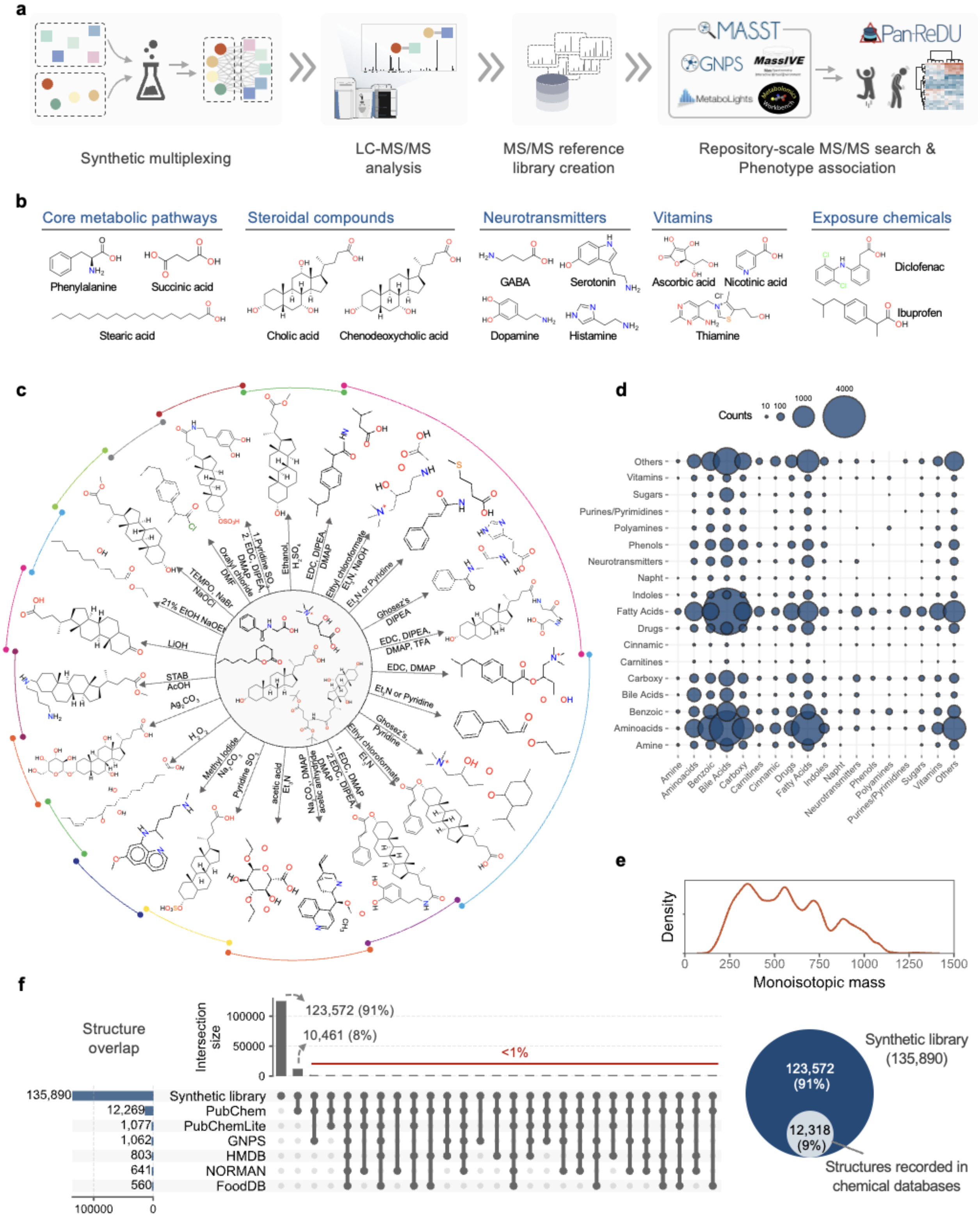

**Figure 1 | The creation of the multiplexed synthetic MS/MS library. a)** Overview of the multiplexed synthesis based reverse metabolomics[4,5] performed in this work. Some of the products of reactions, such as acyl-chloride formation, result in intermediate reagents that subsequently undergo additional reactions. **b**) Representative molecules used as reagents in the multiplexed reactions. **c**) The types of reactions carried out in multiplexed reactions **d**) Representation of unique structures present in the synthesis output (n=42,697). Chemical class categories include related molecules and derivatives. **e**) Mass distribution of the compounds that are part of the MS/MS library. **f**) Evaluation of the uniqueness of the MS/MS library compared to other structural databases.

Using the indexed fast MASST implementation, we searched the newly created MS/MS reference library against the public datasets (**Fig. 2a**). Fast MASST was performed using ≥0.7 cosine score and ≥4 matching ions, criteria that typically result in an FDR <1%[21]. This search yielded matches to 60,146,352 indexed MS/MS spectra in pan-repository data. When combined with existing GNPS reference MS/MS libraries, a total of 8.1% of all MS/MS spectra across the indexed data across the repositories now have a library match, corresponding to a total annotation growth of 17.4%. Both the multiplexed synthesis library and existing libraries provide an initial structural hypothesis when a match within user-defined scoring criteria is obtained. Across all datasets, 63,369 MS/MS spectra from the multiplexed synthesis library were matched, corresponding to 15,190 distinct candidate structures (**Fig. 2b**). UMAP-based visualization of presence/absence patterns of the MS/MS across each taxonomic levels revealed that the MS/MS of many molecules were broadly distributed across many orders and other taxonomic levels (e.g., plants, fungi, animals), suggesting core or possibly even part of yet-to-be documented central metabolism, while others appeared to be taxon-specific (**Fig. 2c-d, Supplementary Figures 1-6**). MicrobeMASST[12]–which enables MS/MS searches against ~60,000 LC-MS/MS of taxonomically defined microbial monocultures–revealed that 24,997 MS/MS spectra of synthesized compounds matched. After removing any candidate compounds that also matched to cultured human cells, this represents 4,596 candidate structures, or some related structural isomer, of putative microbial origin. Based on taxonomic information, most of the MS/MS matched to cultured data from the bacterial phyla belonging to Actinomycetota, Pseudomonodata, Bacteroidota, and to a lesser degree Bacillota (**Fig. 2e**). In addition, we see matches to different fungi such as the Ascomycota and Basidiomycota phyla (**Fig. 2e**). This highlights that there is a large number of microbial molecules that can be readily accessed through synthesis that await to be explored. It should however be noted that the prevalence of the frequency is biased by the number of samples and conditions of LC-MS/MS files for a given taxonomic assignment available in the public domain. Based on NPClassifier[22], molecules derived from alkaloids, fatty acids and terpenoids had its largest share of matches.

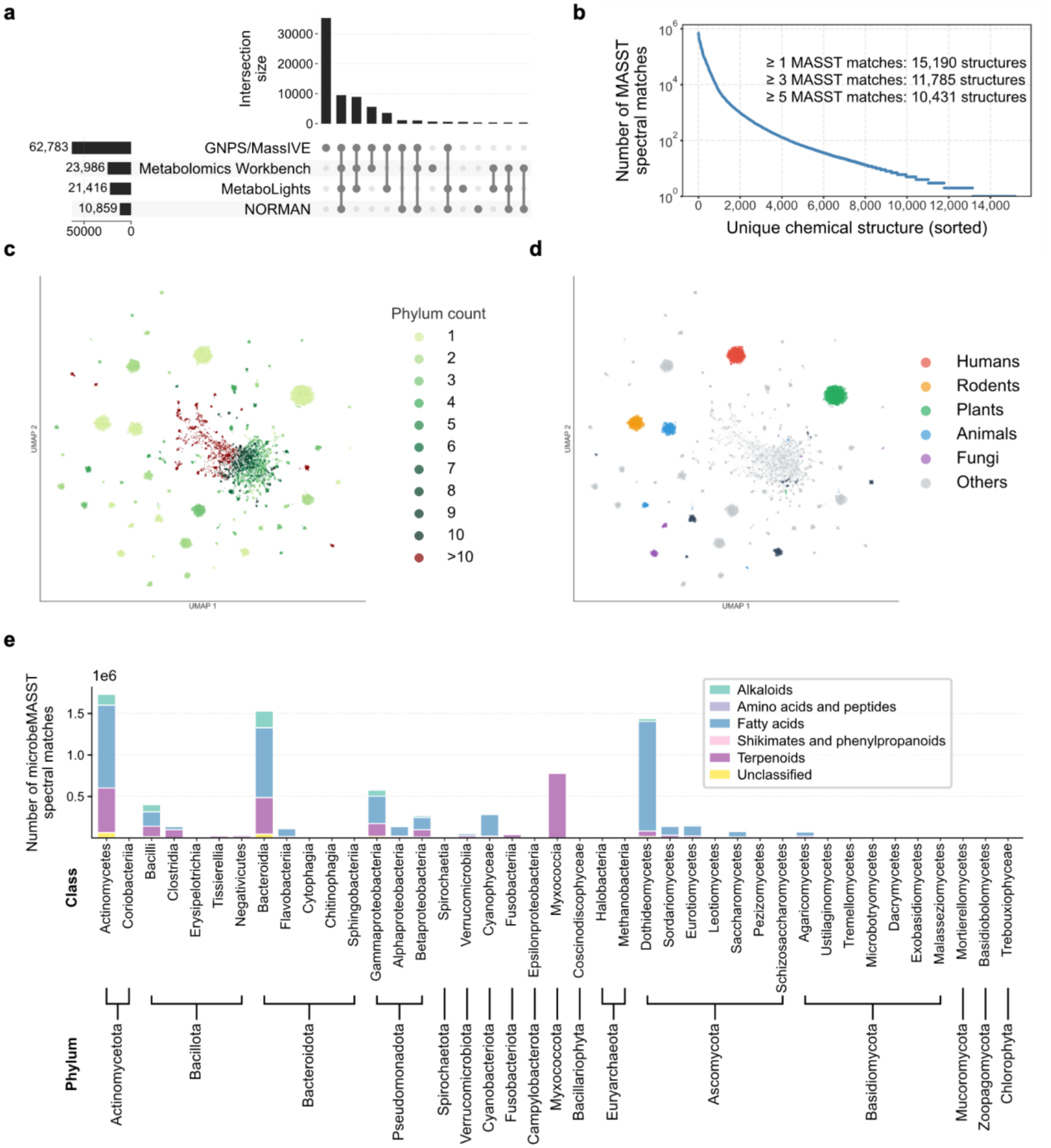

**Figure 2 | Large-scale reverse metabolomics across public datasets using the multiplexed synthetic MS/MS library. a)** Upset plot showing the number of unique MS/MS spectra matched from the synthetic library to public datasets in GNPS/MassIVE, MetaboLights, Metabolomics Workbench, and NORMAN. **b)** Number of MASST spectral matches of the synthetic library for unique chemical structures. **c-d)** UMAP visualization of the taxonomic composition for a given MS/MS spectrum from the multiplexed library at the order level to which we had matched the MS/MS spectra across public datasets, highlighting both taxon-specific and widely shared metabolite signals. Each dot in the UMAP is a MS/MS spectrum from the multiplex library. UMAP of other taxonomic levels can be found in **Supplementary Figures 1-6. e)** Number of microbeMASST spectral matches of the synthetic library across microbial classes and phyla.

Of the 27,807 MS/MS spectra from the multiplexed synthetic library that matched *Homo sapiens* datasets, 2,679 were exclusive to human data (**Fig. 2d**), representing 1,404 candidate structures. Of these spectra, 6.0% (n=161) were derivatives of drug molecules. Examples include derivatives of ibuprofen, 5-ASA, atorvastatin, atenolol, primaquine, naproxen and methocarbamol (**Supplementary Table 4**). Others include bile acids and their derivatives, fatty amides, peptides, carbohydrates, polyketides, shikimites, phenylpropanoids and alkaloid molecules. That we see matches to MS/MS generated from multiplexed reactions with human drugs to human data only makes sense as, generally, other organisms (animals, including rodents, microbes and plants) are generally not given these specific pharmaceutical compounds in the experiments that led to the generation of the untargeted metabolomics data available in the public domain. These spectra associated with humans were distributed across multiple body sites (**Fig. 3a**), with fecal samples showing the highest prevalence. Molecular networking of compounds detected exclusively in human samples revealed candidate drug-related metabolites, including MS/MS matches to 56 and 41 ibuprofen and of 5-aminosalicylic acid (5-ASA) conjugates, respectively (**Fig. 3b**), the majority of which have not been previously reported. We obtained MS/MS matches corresponding to 29 5-ASA derivatives and 33 ibuprofen derivatives across 453,005 human LC-MS/MS datasets in Pan-ReDU (September 2025, **Fig. 3c-h**). MS/MS matches corresponding to 5-ASA derivatives were predominantly detected in human fecal datasets, whereas ibuprofen conjugate spectra were most frequently observed in human urine (**Fig. 3c,d**). Representative MS/MS matches are shown in **Fig 3e-h** and all others can be found as **Supplementary Figure 9a-h**).

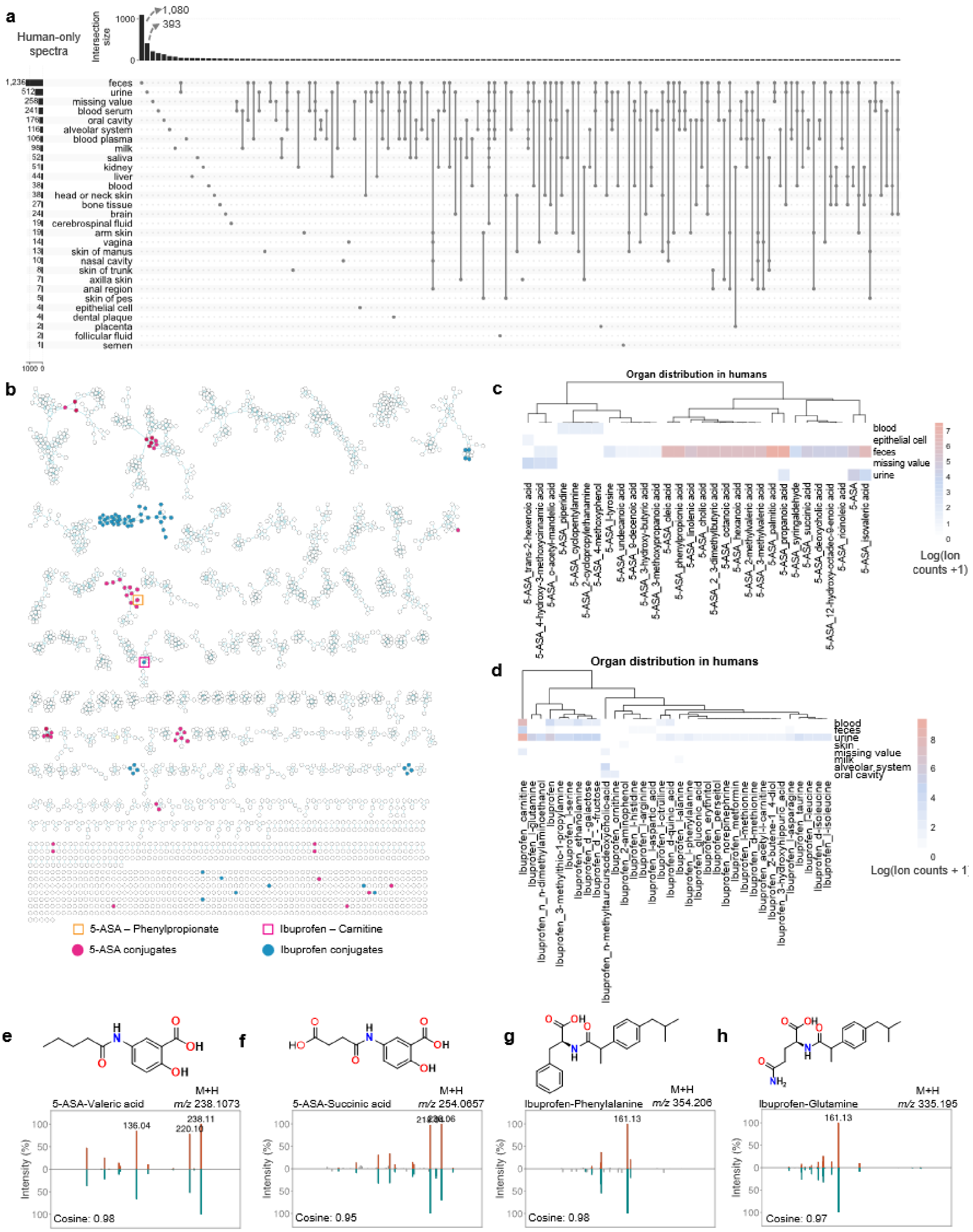
a
Human-only spectra
Intersection size
1000
0
1,080
393
1,236 feces
512 urine
258 missing value
241 blood serum
176 oral cavity
116 alveolar system
106 blood plasma
98 milk
52 saliva
51 kidney
44 liver
38 blood
38 head or neck skin
27 bone tissue
24 brain
19 cerebrospinal fluid
19 arm skin
14 vagina
13 skin of manus
10 nasal cavity
8 skin of trunk
7 axilla skin
7 anal region
5 skin of pes
4 epithelial cell
4 dental plaque
2 placenta
2 follicular fluid
1 semen
1000 0
b
5-ASA – Phenylpropionate
Ibuprofen – Carnitine
5-ASA conjugates
Ibuprofen conjugates
c
Organ distribution in humans
blood
epithelial cell
feces
missing value
urine
5-ASA_trans-2-hexenoic acid
5-ASA_4-hydroxy-3-methoxycinnamic acid
5-ASA_o-acetyl-mandelic acid
5-ASA_piperidine
5-ASA_cyclopentylamine
5-ASA_2-cyclopropylethanamine
5-ASA_4-methoxyphenol
5-ASA_l-tyrosine
5-ASA_undecanoic acid
5-ASA_9-decenoic acid
5-ASA_3-hydroxy-butyric acid
5-ASA_3-methoxypropanoic acid
5-ASA_oleic acid
5-ASA_phenylpropionic acid
5-ASA_linolenic acid
5-ASA_cholic acid
5-ASA_2_3-dimethylbutyric acid
5-ASA_octanoic acid
5-ASA_hexanoic acid
5-ASA_2-methylvaleric acid
5-ASA_3-methylvaleric acid
5-ASA_palmitic acid
5-ASA_propanoic acid
5-ASA_syringaldehyde
5-ASA_succinic acid
5-ASA_deoxycholic acid
5-ASA_12-hydroxy-octadec-9-enoic acid
5-ASA_ricinoleic acid
5-ASA
5-ASA_isovaleric acid
7
6
5
4
3
2
1
0
Log(Ion counts +1)
d
Organ distribution in humans
blood
feces
urine
skin
missing value
milk
alveolar system
oral cavity
Ibuprofen_carnitine
Ibuprofen_l-glutamine
Ibuprofen_n_n-dimethylaminoethanol
Ibuprofen
Ibuprofen_3-methylthio-1-propylamine
Ibuprofen_l-serine
Ibuprofen_ethanolamine
Ibuprofen_d_-galactose
Ibuprofen_d_-fructose
Ibuprofen_n-methyltaurodeoxycholic acid
Ibuprofen_ornithine
Ibuprofen_2-aminophenol
Ibuprofen_l-histidine
Ibuprofen_l-arginine
Ibuprofen_l-aspartic_acid
Ibuprofen_l-citrulline
Ibuprofen_d-quinic_acid
Ibuprofen_l-alanine
Ibuprofen_l-phenylalanine
Ibuprofen_gluconic_acid
Ibuprofen_erythritol
Ibuprofen_perseitol
Ibuprofen_norepinephrine
Ibuprofen_metformin
Ibuprofen_l-methionine
Ibuprofen_d-methionine
Ibuprofen_acetyl-l-carnitine
Ibuprofen_2-butene-1_4-diol
Ibuprofen_3-hydroxyhippuric_acid
Ibuprofen_l-asparagine
Ibuprofen_taurine
Ibuprofen_l-leucine
Ibuprofen_d-isoleucine
Ibuprofen_l-isoleucine
8
6
4
2
0
Log(Ion counts + 1)
e
H
N
O
OH
O
OH
5-ASA-Valeric acid
M+H
m/z 238.1073
136.04
238.11
220.10
Intensity (%)
100
50
0
50
100
Cosine: 0.98
f
HO
O
H
N
O
O
OH
OH
5-ASA-Succinic acid
M+H
m/z 254.0657
218.06
236.06
Intensity (%)
100
50
0
50
100
Cosine: 0.95
g
O
OH
O
N
H
Ibuprofen-Phenylalanine
M+H
m/z 354.206
161.13
Intensity (%)
100
50
0
50
100
Cosine: 0.98
h
O
OH
O
N
H
O
NH2
Ibuprofen-Glutamine
M+H
m/z 335.195
161.13
Intensity (%)
100
50
0
50
100
Cosine: 0.97

**Figure 3 | Molecules from the synthetic MS/MS library matched to human only data with MASST. a)** UpSet plot showing the number of matched compounds associated with each parent drug and their overlaps across drugs. **b)** Molecular network of all matched compounds, with each parent drug and its associated matched derivatives colored distinctly. Two example clusters are highlighted: 5-aminosalicylic acid (5-ASA) and ibuprofen. **c–d)** Organ-level distribution of all matched derivatives for (**c**) 5-ASA and (**d**) ibuprofen across available human datasets, showing where these compounds were detected. **e–h)** Representative MS/MS mirror plots and chemical structures for selected matched analogs of 5-ASA and ibuprofen. Each plot displays the experimental MS/MS spectrum from the synthetic library (top) and the matched human spectrum (bottom), along with the corresponding compound structure. Full mirror plots and structures for all 5-ASA and ibuprofen derivatives are available in the Supplementary Information.

Although we used typical scoring conditions of cosine of 0.7 or higher that typically lead to less than 1% FDR for MS/MS spectral alignment[21], MS/MS matches to reference libraries are always considered a structural hypothesis rather than a confirmed structural entity. We therefore set out to provide additional experimental validation of the existence of the 5-ASA and ibuprofen derived annotations. In order to provide additional confirmation of the existence of these drug derived metabolites we would need to validate them with retention time and/or drift time in human samples. To accomplish this, given that our matches thus far were limited to public datasets, we used MASST to identify public-domain samples containing MS/MS spectra matching 5-ASA conjugates we could match our standards against. Matches were found from inflammatory bowel disease (IBD) fecal datasets that we were able to get access to, enabling confirmation with MS/MS, retention time, and ion mobility using two different LC-MS/MS instruments (**Fig. 4a-c**). Across these fecal samples, we matched seven 5-ASA conjugates against synthetic standards, including four long-chain fatty acid conjugates, two short-chain fatty acid conjugates, and a phenylpropionate conjugate (**Fig. 4b**). Short-chain fatty acid-derived 5-ASA metabolites have been previously reported as microbial metabolism products[23,24], whereas the remaining conjugates have not been described before.

Focusing on the phenylpropionic acid, a known microbial metabolite[25], we detected both 5-ASA and its phenylpropionic acid conjugate in feces (**Fig. 4c,d**). This represents a previously unreported microbiome-derived 5-ASA metabolite, likely formed via microbiome-host or microbe-microbe co-metabolism. This conjugate was quantified in a fecal sample to be 20.8 µM (quantification details are available in the methods). Using MicrobiomeMASST, a tool in development, which links metabolites to microbiome-relevant information such as organs, age, interventions, and health conditions, 5-ASA-phenylpropionate was detected in 41/693 data files labeled as IBD, 6/333 data files labeled with rheumatoid arthritis (RA), and 1/14,567 healthy human samples. Compared to data from healthy individuals samples, detection was higher in IBD (odds ratio [OR] = 916, 95% CI = 126–6669, Fisher's exact p = 1.1 × $10^{-54}$) and RA (OR = 267, 95% CI = 32–2226, p = 8.2 × $10^{-10}$). Direct comparison between IBD and RA revealed higher odds in IBD (OR = 3.43, 95% CI = 1.44–8.16, p = 0.0023), consistent with the more widespread clinical use of 5-ASA in IBD compared to RA. The single positive control labeled as healthy we hypothesize was labeled incorrectly in the public domain data. It was part of a contrasting Western group in a microbiome study of remote villages[26] and we hypothesized that this sample information to have been misassigned as clinical status was likely assumed to be healthy by the data depositor as it was a control group for a non-clinical study. Its inclusion therefore provides conservative effect size estimates, as exclusion would only further increase the odds ratios and strengthen the associations. Thus, while wide confidence intervals reflect uncertainty due to the rarity of the 5-ASA-phenylpropionate, enrichment in IBD and RA is robust, and the true effect is likely underestimated in our reported values.

As 5-ASA treatment is not universal among patients with IBD or RA, we next compared the observed metabolite detections to clinical metadata documenting 5-ASA or related prodrug administration, where available (**Supplementary Figure 7**). In three studies - two IBD cohorts and one RA cohort - with documented 5-ASA or prodrug usage (e.g., sulfasalazine, balsalazide), we compared 5-ASA-phenylpropionate matches to medication metadata. In a pediatric IBD cohort, 2 of 6 exposed study participants had matches, while 0 of 46 unexposed individuals did (**Fig 4e**). Combining all three datasets, structure matches were observed in 9/29 exposed and 1/83 of non-exposed had detection. This supports they were enriched in exposed individuals compared to unexposed controls (OR = 36.9, 95% CI 4.4 - 308.3, one-tailed Fisher's exact test, $p$ = 9 × $10^{-}5$), further supporting a drug origin for the phenylpropionate-linked 5-ASA compound consistent with known prescription patterns. In addition, it was observed in the monocultures of *Bacteroides caccae*, *Bacteroides vulgatus*, *Bacteroides luhongzhouii, Bacteroides thetaiotaomicron*, and in a 12-member Crohn's diseases synthetic community, to which phenylpropionic acid and 5-ASA were added to the growth culture. Thus, confirming the microbial origin of the drug conjugate **(Supplementary Figure 7a)**.

To confirm that 5-ASA conjugates reflect drug-specific exposure rather than baseline metabolic processes, we compared them to succinylated amines, metabolites expected to occur broadly. Succinic acid, a core intermediate of primary metabolism, reacts with diverse amines and is widely distributed across organisms and health states[27–29]. Indeed, Pan-ReDU yielded MS/MS matches to 26 succinylated amines (**Fig. 4f,g**), spanning humans, microbes, and other organisms (**Supplementary Figure 8a-b**), with a large portion of the human matches (48.5%, n=928/1913) observed in healthy-labeled datasets. In contrast to the disease- and medication-restricted distribution of 5-ASA conjugates, succinylated amines were broadly detected, including in non-human data, highlighting that 5-ASA derivatives serve as specific markers of drug exposure rather than baseline metabolic products.

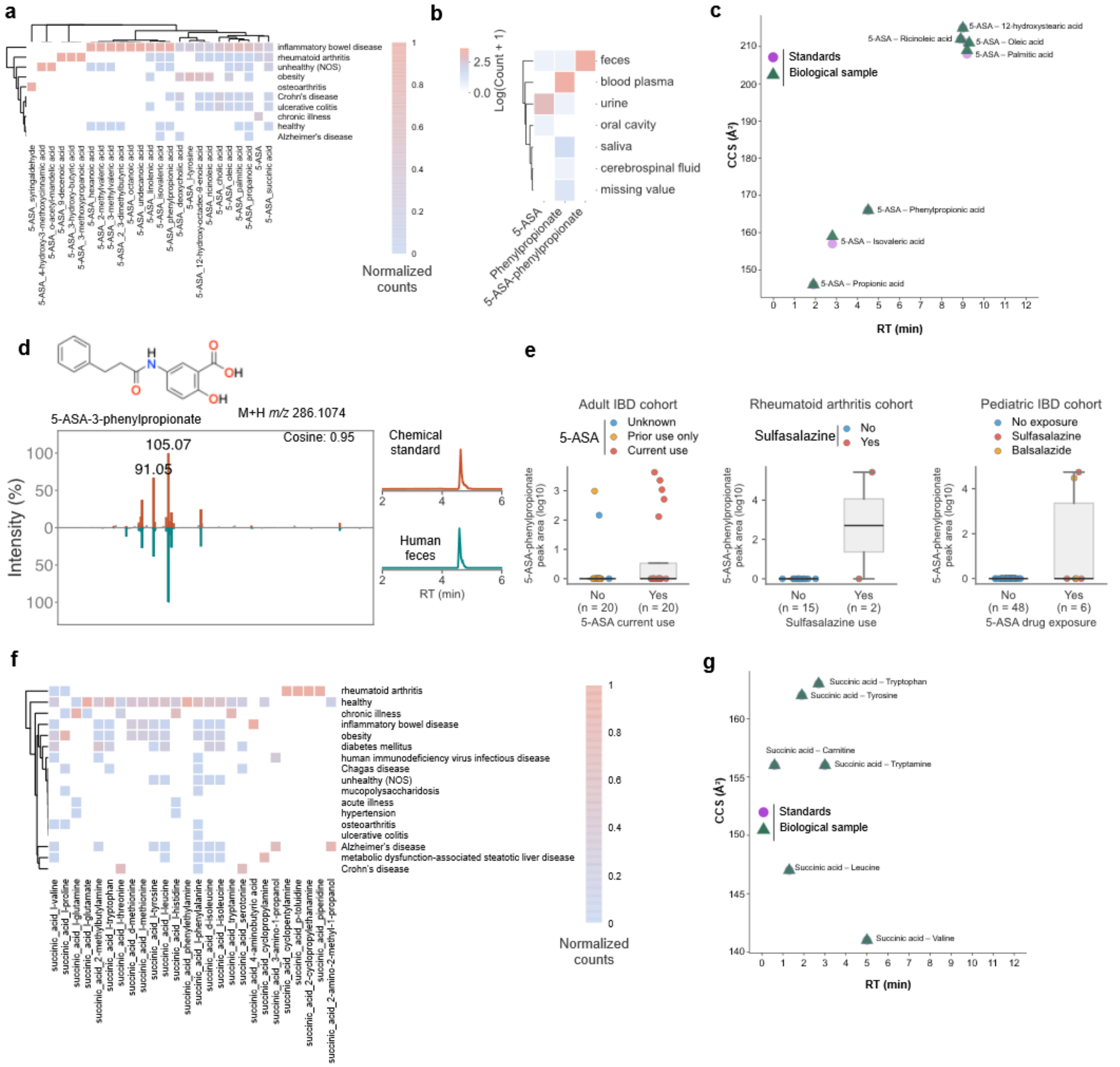

**Figure 4 | Characterization and validation of 5-aminosalicylic acid (5-ASA) derivatives in human data. a)** Association of 5-ASA and its derivatives with specific health conditions, highlighting links to IBD and RA. **b)** Comparative organ-level distribution of 5-ASA-phenylpropionate and its unconjugated counterpart phenylpropionate across human datasets. **c)** Experimental validation of 5-ASA derivatives using retention time (RT) and collision cross section (CCS) measurements, confirming MS/MS-based annotations (full MS/MS spectra in Supplementary Information). **d)** Representative MS/MS mirror plot and RT validation for 5-ASA-phenylpropionate, showing the synthetic spectrum (top) and matched human spectrum (bottom), alongside the compound structure. **e)** Relative intensities of 5-ASA derivatives across two independent IBD and RA studies, illustrating disease-associated differences. **f)** Comparison of succinylated derivatives of 5-ASA across health conditions, demonstrating additional patterns of disease relevance. **g)** CCS and RT matches of succinic acid conjugates to chemical standards.

In contrast to succinylated amines and more similar to 5-ASA derivatives, the MS/MS matches to the 33 candidate ibuprofen-derived conjugates were found exclusively in human datasets, including data from healthy individuals (**Fig. 5a**). The many diabetes matches could be primarily driven by the preponderance of urine data for this health condition that are not as prevalent in other health conditions in the public domain data. Thus, although the biology may warrant further exploration, this observation could also reflect the database composition. Seven of these metabolites were verified by MS/MS and retention time across two different instrument platforms, one of which also provided ion mobility data, from human urine (**Fig. 5b-d**). In contrast to reported phase I/II transformations such as hydroxylation, carboxylation, glucuronidation, and taurine addition, we found no reported evidence for these conjugates in rodents or humans that were exposed to ibuprofen[30–33]. Instead, the ibuprofen-carnitine conjugate was widely observed across human datasets and across health categories **(Supplementary Figure 7b)**, including healthy individuals - consistent with common over the counter use for muscle aches and headaches of individuals who would generally consider themselves healthy. It was detected in 105 human datasets of 61 files with available harmonized sample information, spanning feces, urine and serum/plasma, with urine being the most common matrix (**Fig. 5c**). Quantification in a urine sample gave a concentration of 1.37 µM. The presence of the conjugate in urine prompted us to test its consistency across samples with likely ibuprofen exposure. Analysis of fresh urine from an ongoing urogenital microbiome and metabolomic profiling study was performed to assess if the carnitine metabolite (and other ibuprofen derived conjugates) are present. Urine samples were obtained from hospitalized patients likely to be receiving ibuprofen as part of their clinical care. The ibuprofen–carnitine conjugate was detected in all nine samples that also contained ibuprofen (**Fig. 5e**), and it was the most abundant ibuprofen-derived compound in all but one case, where carboxy-ibuprofen dominated. Importantly, discovery of these NSAID-carnitine conjugates was enabled by the multiplexed MS/MS library, which allowed structural annotation of metabolites that have eluded conventional metabolomics reference libraries which are largely restricted to commercially available compounds rather than previously undescribed metabolites - despite ibuprofen's FDA approval in 1974 and over-the-counter availability since 1984.

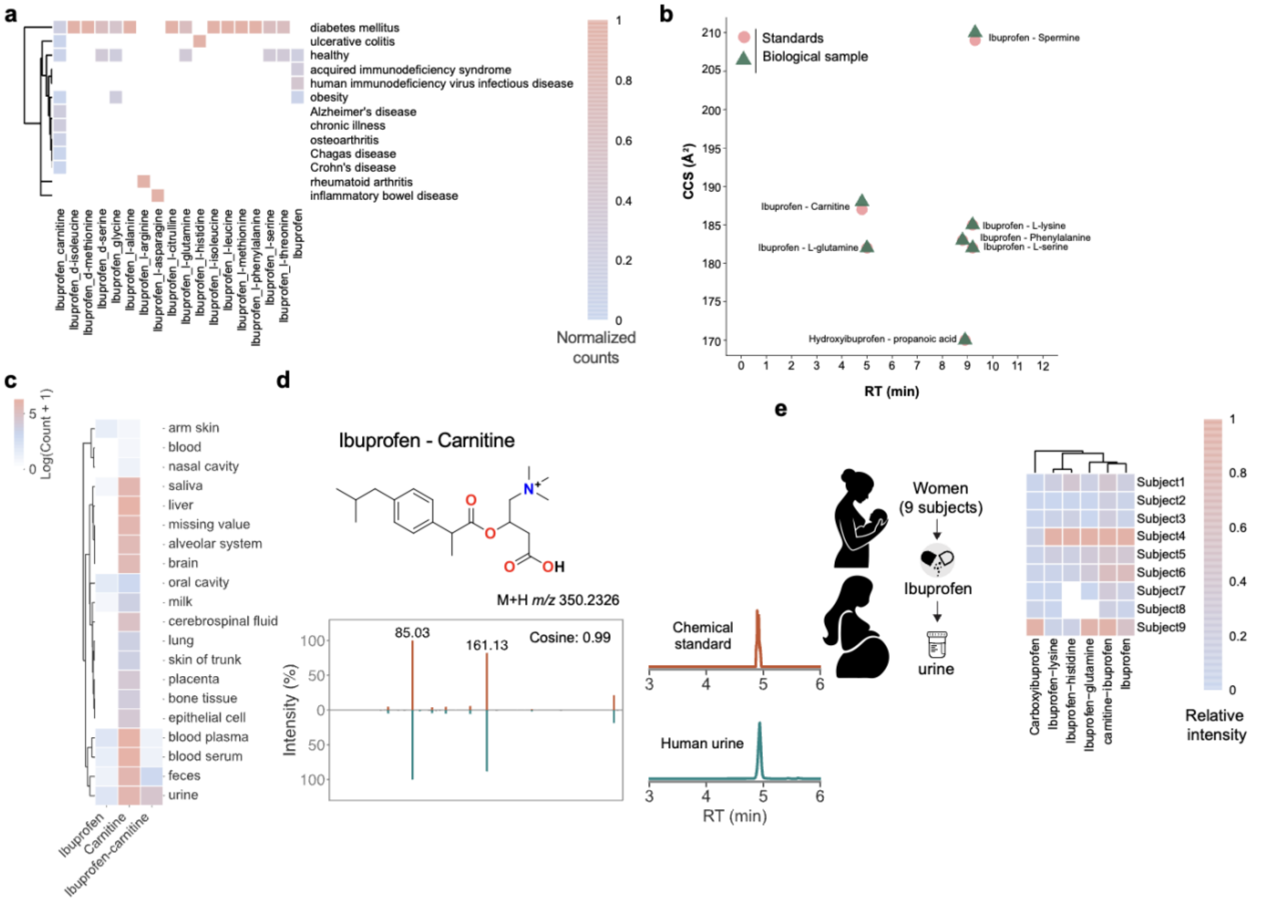

**Figure 5 | Characterization and validation of ibuprofen-derived metabolites in human datasets. a)** Distribution of the MS/MS of the ibuprofen parent compound and its conjugated derivatives across health conditions, highlighting associations with medication exposure and relevant disease groups. Each derivative is labeled by chemical class (e.g., carnitine, glucuronide, sulfate, acyl-conjugates). **b)** Experimental validation of ibuprofen derivatives using retention time (RT) and collision cross section (CCS) measurements; MS/MS spectra supporting annotations are provided in the Supplementary Information (or in the main figure, if included). Reported RT and CCS values are from authentic synthetic standards. **c)** Organ-level distribution map for IB-carnitine (representative ibuprofen-carnitine conjugate) showing repository-wide detection across tissues and biofluids; heatmap/points indicate the presence and relative frequency of matches in each organ dataset. **d)** Representative MS/MS mirror plot and RT validation for Ibuprofen-carnitine: synthetic library spectrum (top) versus matched human spectrum (bottom), with the annotated chemical structure and reported RT/CCS concordance. **e)** Distribution of all detected ibuprofen analogs in a hospitalized cohort demonstrating observed analog classes and relative intensities/frequencies in urine.

Prior to the discovery of ibuprofen-carnitine in humans, some NSAID-carnitine conjugates had been explored synthetically (e.g., naproxen- and ketoprofen-carnitine derivatives but not ibuprofen), motivated by the potential to enhance cellular uptake via the OCTN2 carnitine transporter and improve drug delivery to renal tissues while reducing systemic toxicity[34]. However, their occurrence *in vivo* in humans had not been demonstrated. NSAID-carnitine conjugates may also impact carnitine metabolism and transport, as OCTN2 recognizes naproxen- and ketoprofen-carnitine derivatives as both substrates and inhibitors[35]. Thus, detection of ibuprofen-carnitine in human samples raises the possibility that such metabolites could contribute to NSAID-related adverse effects, including mitochondrial or muscle toxicity, through carnitine depletion or transporter competition - particularly under chronic exposure, high carnitine needs such as muscle recovery in endurance athletes or in genetically susceptible individuals. This hypothesis warrants further investigation.

Together, these results demonstrate that reanalysis of public metabolomics data with a biochemically inspired multiplexed MS/MS library can uncover previously unrecognized metabolites. In total, we detected putative MS/MS matches to 15,190 molecules in the public domain, of which 20 were elevated to level 1 identification according to the Metabolomics Standards Initiative through confirmation with MS/MS, retention time, and with ion mobility matching[10]. These annotations expand the known metabolic map, encompassing products of primary metabolism, host–microbe co-metabolism, and drug biotransformations. Importantly, the contrasting repository-scale distributions of 5-ASA conjugates, ibuprofen conjugates, and succinylated amines highlight how drug-derived metabolites - restricted to human data and reflective of medication use - can be distinguished from more broadly distributed primary metabolic conjugates. As it is not a traditional MS/MS library of compounds that can be purchased, we provide suggestions for its proper use as below.

**It is important to properly use and interpret the library and understand its limitations:** Matches to the multiplexed MS/MS library, like any spectral reference in untargeted metabolomics, should be interpreted as plausible structural hypotheses rather than definitive identifications. Even high-scoring matches can be confounded by stereochemistry, positional or geometric isomerism, adduct variation, and fragmentation ambiguity. Tandem MS alone cannot generally unambiguously resolve single structure, so annotations must be evaluated in the context of biosynthetic logic, sample metadata, and orthogonal validation such as retention time, ion mobility, or isolation and NMR or other additional structural analysis. Assessing biological plausibility provides additional confidence. Harmonized metadata from public repositories enables evaluation across thousands of studies. For example, bile acids are animal-specific and their detection in plant datasets should be treated with caution, whereas very long-chain N-acyl lipids (>C24) are common in plants but rare in animals. Drug conjugates such as ibuprofen-carnitine or 5-ASA-phenylpropionate are observed only in human datasets where exposure is expected. Their absence in unrelated contexts strengthens annotation confidence, while unexpected detections warrant deeper scrutiny. Similarly, co-occurrence of related metabolites provides additional evidence. Ibuprofen–carnitine often appears alongside other ibuprofen conjugates, and 5-ASA-phenylpropionate co-occurs with acetate, butyrate, and longer-chain fatty acid or amino acid 5-ASA conjugates, consistent with microbiome-mediated or host co-metabolism. In contrast, isolated detections may indicate rare transformations, false matches, or incomplete sampling. Integrating spectral evidence with co-occurrence and biochemical context helps translate MS/MS similarity into biologically meaningful hypotheses.

Tandem MS cannot reliably distinguish structural isomers based solely on fragmentation. Molecules with identical elemental composition, including those formed by acylation, amidation, or esterification, can yield highly similar spectra. For example, monoacetylation of OH's of cholic acid produces three positional isomers whose MS/MS spectra are nearly indistinguishable under standard collision-induced dissociation. The multiplexed library prioritizes chemical diversity over site specificity, so many entries could represent mixtures or multiple isomers. Unambiguous structural assignment requires orthogonal validation, such as using retention time and drift time, as demonstrated for ibuprofen-, 5-ASA-, and carnitine-derived conjugates. Computational tools such as ICEBERG[36] and Modifinder[37] are expected to enhance isomer discrimination and annotation coverage in the future.

All spectra were acquired on a single instrument platform under defined collision-induced dissociation conditions, detecting primarily $[M+H]^+$, $[M+Na]^+$, and $[M+NH_4]^+$ adducts. Consequently, less common ion forms, multimers, or side products are underrepresented. The library spans over 4,000 reactions collected in positive ion mode, reflecting the majority of public LC-MS/MS data, though negative mode and additional instrument platforms would expand coverage. From ~10 million spectra generated, ~0.5 million were curated into the final library. The remaining spectra likely include uncharacterized analogs, delta-mass derivatives, or in-source fragments. These data are publicly accessible for future reanalysis, reaction discovery, and iterative library expansion through molecular networking and annotation propagation.

Naming the compounds is a major challenge and welcome anyone reading this to reach out for practical solutions. Many synthesized compounds are absent from structural databases, so conventional names are often impractical. IUPAC chemical names generated from SMILES are too complex for routine use. To improve clarity and computational accessibility, a reagent-based naming convention was adopted. For instance, we use the name "Erythro-aleuritic acid_glycine (known isomers: 0; isobaric peaks: 2)" This denotes a reaction between Erythro-aleuritic acid and glycine, with the underscore separating reagents and parentheses indicating predicted isomers and observed peaks. Each entry links to its raw file and one of the possible SMILES representations. While this system enhances traceability, it does not resolve inherent structural ambiguity. Matches should therefore be treated as biologically plausible leads, guiding hypothesis-driven synthesis and annotation refinement through orthogonal validation as done with 5-ASA and ibuprofen conjugates.

## Conclusion

This work establishes a scalable, hypothesis-driven framework for reverse metabolomics by combining biologically inspired multiplexed synthesis, high-resolution MS/MS, and systematic mining of public datasets. The resulting synthetic MS/MS reference library - comprising nearly half a million curated spectra across structurally diverse small molecules - enables broad exploration of previously unannotated chemical space. Through iterative workflows of match → hypothesis → synthesis → reanalysis, researchers can uncover unexpected biochemical transformations, such as those demonstrated with bile acids, N-acyl lipids, carnitine, carbohydrates and clinically relevant drug conjugates with ibuprofen and 5-ASA.

This approach is not static: it will evolve. A key long-term goal is to curate every MS/MS spectrum - irrespective of ion form, adduct, or fragmentation condition - so that each signal in LC-MS/MS based metabolomics data can eventually be linked to an interpretable structural hypothesis. Even when multiple ion forms or in-source fragments get annotated, their inclusion enables researchers to make informed decisions about how to process, quantify, or exclude those signals depending on their biological relevance and analytical context. As new hypotheses arise and uncharacterized MS/MS features accumulate, the system can be expanded to include additional compound classes - ranging from dietary and environmental exposures to microbiome- and host-derived metabolites. Future efforts should also incorporate more complex, multi-step, or enzyme based synthetic transformations to further mimic biochemical metabolism and extend annotation capacity into deeper regions of chemical space that are not yet being explored.

While tandem MS has inherent limitations - including difficulty distinguishing structural isomers, variability in ion forms - these challenges are met with scalable data science strategies. Molecular networking and nearest-neighbor propagation allow for class-level annotations beyond exact matches. Mass difference analysis would be expected to link a significant portion of unmatched features to related molecules[38] that did not yet make it in the 2025 multiplexed library. These would correspond to predictable modifications of curated structures, emphasizing the opportunity to grow a future library, leveraging this same data. For such expansion understanding the reagents that were used would further enhance structurally informed propagation.

Ultimately, this multiplexed synthesis strategy represents a unique route to illuminate the “dark matter” of the metabolome[2]. It facilitates data-driven structural hypothesis generation, structural anchoring of unknowns, and scalable annotation workflows that bridge synthetic chemistry, informatics, and biology. In doing so, it will also require shifts in the field from static pathway representations of hand curated metabolic pathway maps toward dynamic, interconnected computationally created metabolic networks that not only can handle annotation ambiguity but also reflect the true diversity and complexity of life’s chemistry - empowering future discoveries across metabolomics, exposomics, pharmacology, and systems biology.

## Methods

We have significantly expanded the reverse metabolomics approach, which associates MS/MS spectral profiles of the synthesized compounds with biological phenotypes through analysis of extensive public untargeted metabolomics repositories. We have extended its capacity in both computational capabilities and by creating a large reference spectral library to enable more comprehensive discovery of complex metabolites and its chemical–biological association. At its foundation, reverse metabolomics identifies the occurrence of any submitted MS/MS spectrum within public repositories and subsequently utilizes the associated metadata to correlate metabolites with a range of experimental variables, including disease states, taxonomic distribution, and sample types.

To validate the enhanced scalability of this methodology, we generated a library of structurally diverse candidate metabolites via multiplex synthesis and acquired corresponding LC-MS/MS data. Our multiplex synthesis encompassed five principal compound classes: amino acid conjugates (N-acyl amides, glutathione adducts, and peptide derivatives); microbial–host co-metabolites (bile acid amidates and bile acid esters), secondary metabolites (phenolic glycosides, alkaloids, and terpenoids); lipid classes and derivatives (fatty acid esters, glycerophospholipids, and sphingolipids); and xenobiotics (drug metabolites, environmental contaminants, and dietary compounds). The occurrence of the synthesized compounds in the public domain was obtained using Mass Spectrometry Search Tool (MASST)[7], and the relevant metadata was analyzed and assessed using Reanalysis of Data User Interface (Pan-ReDU)[11].

The multiplex synthetic MS/MS spectra exhibit a distribution of molecular ion forms, including 297,483 spectra (60.4%) corresponding to $[M+H]^+$, 92,872 spectra (18.9%) to $[M+NH_4]^+$, and 53,269 spectra (10.8%) to dehydrated ions ($[M+H–H_2O]^+$). The remaining spectra represent less common adducts, such as $[M+Na]^+$, $[M+K]^+$, and doubly dehydrated ions ($[M+H–2H_2O]^+$).

## Chemical class prediction

The compound class information of newly synthesized chemicals were predicted using NPClassifier[22], a deep neural network-based tool for structural classification. This was programmatically achieved via the GNPS2 API using the SMILES strings. For compounds which had more than one possible compound pathways provided by NPClassifier, the first pathway was reserved for downstream analysis and visualization.

## Sample collection and extraction

Urine samples were collected as part of an ongoing prospective cohort study for benchmarking storage and processing of the urogenital microbiome and metabolomic profiles (UC San Diego IRB#801735). Urine samples were obtained from hospitalized patients likely to be receiving ibuprofen as part of their clinical care. Voided urine was self-collected by participants and aliquoted and frozen at -80 until extraction. Urine samples were prepared as previously described[39]. In brief, a 200 µL aliquot of urine was transferred into an empty sample tube. Then, 800 µL of 80% methanol was added, resulting in a final volume of 1 mL. Samples were vortexed for 5 s and then incubated at −20 °C for 20 min for protein precipitation. Following incubation, samples were centrifuged at 2000 rpm for 5 min at 4 °C to pellet the precipitated proteins. A volume of 800 µL of the resulting supernatant was transferred into the wells of a pre-labeled 96 deep-well plate. Samples were dried using a centrifugal vacuum concentrator (Centrivap). The dried residues were reconstituted in 250 µL of a 50% methanol-water solution containing 1 µM sulfadimethoxine as an internal standard. Fecal samples were obtained from a pilot study investigating the influence of diet on patients with rheumatoid arthritis (UC San Diego IRB#161474). Stool samples were weighted and extracted at a ratio of 50 mg of sample to 800 µL of 50% MeOH/$H_2O$. A 5 mm stainless steel bead was added to the samples and homogenized using a Qiagen TissueLyser II for 5 min at 25 Hz, before being incubated overnight at 4 °C. Samples were centrifuged at 15,000 x g, 200 µL was transferred, and dried using a CentriVap. All samples were resuspended with 200 µL containing an internal standard and incubated at -20 °C overnight. All samples were centrifuged at 15,000 x g and 150 µL of supernatant was transferred into a glass vial for LC-MS/MS analysis.

**LC-MS/MS data collection**

Biological samples and the synthetic standards were obtained for retention time and MS/MS spectral matching and were subjected to LC-MS/MS analyses. The LC-MS/MS analyses were carried out with a Vanquish UHPLC system coupled to a Q-Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The chromatographic separation was performed on a Polar C18 column (Kinetex C18, 100 x 2.1 mm, 2.6 µm particle size, 100A pore size – Phenomenex, Torrance, USA), and the mobile phase consisted of $H_2O$ (solvent A), and ACN (solvent B), both acidified with 0.1% formic acid. The following gradient was employed to evaluate retention time matching between synthetic standards and the compounds present in the samples: 0-0.5 min 5% B, 0.5-1.1 min 5-20% B, 1.1-5.0 min 20-40% B, 5.0-9.0 min 40-100% followed by a 1.5 min washout phase at 100% B, and a 1.5 min re-equilibration phase at 5% B. The flow rate was set at 0.5 mL/min, the injection volume was fixed at 3 µL, and the column temperature was set at 40 °C. Data-dependent acquisition (DDA) of MS/MS spectra was performed in the positive ionization mode. Electrospray ionization (ESI) parameters were set as: 52.5 AU sheath gas flow, 13.75 AU auxiliary gas flow, 2.7 AU spare gas flow, and 400 °C auxiliary gas temperature; the spray voltage was set to 3.5 kV and the inlet capillary to 320°C and 50 V S-lens level was applied. MS scan range was set to 300-800 *m/z* with a resolution of 35,000 with one micro-scan. The maximum ion injection time was set to 100 ms with an automated gain control (AGC) target of 1.0E6. Up to 5 MS/MS spectra per MS1 survey scan were recorded in DDA mode with a resolution of 17,500 with one micro-scan. The maximum ion injection time for MS/MS scans was set to 150 ms with an AGC target of 5E5 ions. The MS/MS precursor isolation window was set to 1 *m/z* with an offset of 0 *m/z*. The normalized collision energy was set to a stepwise increase from 25, 40, and 60 with $z = 1$ as the default charge state. MS/MS scans were triggered at the apex of chromatographic peaks within 2 to 5 s from their first occurrence. The quality and reproducibility of the analyses were evaluated considering the retention time and the *m/z* of a standard solution containing a mixture of six standards (amytriptiline, sulfamethizole, sulfamethoxine, sulfadimethoxine, coumarin 314, and sulfachlopyridazine) which was analyzed every five samples.

**MS/MS spectral library generation**

**Structure generation of expected product molecules**

To generate the structural information (SMILES strings) of the expected chemical products, we first create an input csv file containing compound names and SMILES strings of the reactants. This file is then uploaded to the AutoSMILES app. Next, we select which reactant to use as the sample ID, and specify the first and second reactant locations in the input csv file. Then, we enter the desired number of decimal places for mass precision. The type of reaction to perform is then selected - for example, amidation, esterification, hydroxylation, methylation, etc. You can also enter any output file filler values that you want to include as additional columns in the output .csv file. Finally, click Start Reaction, and the app will generate all the product SMILES strings for the input reactants.The pipeline can be accessed via https://autosmiles.streamlit.app/rxnSMILES.

### MS/MS spectra retrieval from raw LC-MS/MS data

Raw LC-MS/MS data collected for the multiplex synthesis library creation were first converted into an open format (mzML) using MSConvert. Then the mzML files and the csv files containing product SMILES strings were uploaded to the GNPS2 (https://gnps2.org/homepage) file browser. The reverse_metabolomics_create_library_workflow was applied on the input mzML files and compound csv files, creating the MS/MS library in the format of mgf and tsv. These output mgf and tsv files were then uploaded to the GNPS library https://external.gnps2.org/gnpslibrary. The library generation workflow is available on GNPS2 through https://gnps2.org/workflowinput?workflowname=reverse_metabolomics_create_library_workflow.

### Repository-scale MASST searches

A minimum of 0.7 cosine score, a precursor and fragment mass tolerance of 0.05 DA was used to collect similar or identical MS/MS spectra for the four main metabolomics repositories (GNPS/MassIVE, Metabolomics Workbench, Metabolights, and NORMAN). Any spectral match against the multiplex synthetic datasets deposited on GNPS/MassIVE were removed before analysis. We further filtered the collected spectra by using a minimum of 4 matched peaks.

### Analysis of drug conjugates

Ibuprofen and 5-ASA conjugates were multiplex-synthesized and subjected to the reverse metabolomics workflow[4]. For the MASST analysis as shown in **Fig. 3d**, a minimum of 3 matched peaks was applied to account for small molecules such as 5-ASA and phenylpropionate which do not usually generate more than 3 peaks in their MS/MS spectra. As these drugs were only provided to humans, we removed any drug conjugates with MS/MS that matched to non-human samples, as shown in heatmaps in **Fig. 4b**. A total of 21 ibuprofen and 29 5-ASA conjugates were successfully identified. In addition to these, further matches were observed with other nonsteroidal anti-inflammatory drugs (NSAIDs), including aspirin, naproxen, and various NSAID metabolites. Notably, several matches were also detected with non-NSAID pharmaceuticals, such as atorvastatin, atenolol, metoclopramide, and primaquine, suggesting a broader interaction profile across multiple drug classes.

### Analysis of public GNPS/MassIVE datasets

The multiplex synthetic library created was used to analyze multiple datasets: (1) inflammatory bowel disease (MSV000082094, fecal samples); (2) a pediatric IBD cohort (MSV000097610, fecal samples); (3) a rheumatoid arthritis cohort (MSV000084556, fecal samples). Each public dataset was downloaded and processed using MZmine using the batch workflow for feature finding and detection. An example of a .mzbatch file containing detailed parameter settings can be found in https://github.com/VCLamoureux/synthesis_multiplex. The output files generated using the batch processing workflow (a csv file with peak areas and an mgf file with MS/MS information associated with each feature) were used as input for feature-based molecular networking (FBMN) in GNPS2 and ran against the multiplex synthetic library. The FBMN parameters were set for each dataset with a cosine of 0.7, precursor and fragment ion tolerance of 0.02 Da, filters set to off, and a number of matched peaks for networking and library search, set to 4. For each dataset, the quantification table (generated via MZmine), the annotation table (FBMN workflow), and the metadata was imported in RStudio for data formatting. The formatted data tables were exported into a csv file for boxplots creation using Python scripts (see Code availability).

## Computational infrastructure

This is an expanded description of the computational infrastructure that was developed to enable this project. MASST[7] (Mass Spectrometry Search Tool) queries now run on a virtual machine equipped with two 64-core AMD Milan EPYC 7713 processors and 2 TB of RAM, with public metabolomics data indices hosted on four NVMe Solidigm D5-P5336 SSDs configured in a RAID ZFS striped array. We refer to these as fast MASST or FASST[40] queries. The GNPS2 platform has expanded to operate across five virtual machine servers: two with dual 64-core AMD Milan EPYC 7713 processors and 2 TB of RAM, and three with dual 16-core Intel E5-2683 v4 CPUs and 768 GB of RAM. Storage is provided by two arrays: one comprising 24 × 7.68 TB SATA SSDs (184 TB) and another with 8 × 30 TB NVMe SSDs (240 TB). All servers are interconnected via a 10 Gbit network.

## Materials availability

All the reagents in this study were included in the key **resources** table (key Resources) provided as supplementary information. While we encourage other labs to synthesize the compounds as needed - we will make the reactions from the multiplexed reactions available while supplies remain.

## Chemical synthesis

NMR spectra were collected at 298 K on a 500 MHz Bruker Avance III spectrometer fitted with a 1.7 mm triple resonance cryoprobe with z-axis gradients. ($^1$H NMR: MeOD (3.31), $CDCl_3$ (7.26) at 500 MHz. 5-ASA-phenylpropionic acid spectra was taken in MeOD with shifts reported in parts per million (ppm) referenced to the proton of the solvent (3.31), and Ibuprofen-carnitine spectra was taken in $CDCl_3$ with shifts reported in parts per million (ppm) referenced to the proton of the solvent (7.26), Coupling constants are reported in Hertz (Hz). Data for $^1$H-NMR are reported as follows: chemical shift (ppm, reference to protium; s = single, d = doublet, t = triplet, q = quartet, dd = doublet of doublets, m = multiplet, coupling constant (Hz), and integration).

## Multiplex reactions

The synthesis procedures for 5-ASA-phenylpropionic acid and ibuprofen-carnitine are detailed below. Additionally, the complete set of multiplex library reaction protocols, including all reagents, and conditions are provided in Supplementary Information.

## 5-ASA-phenylpropionic acid

Solid 5-aminosalicylic acid (2 mmol, 100 mg, 1 eq.) and 3 mL of THF were added sequentially to a 20 mL glass vial with a stir bar and the reaction was placed in an ice/water bath, neat triethylamine ($Et_3N$) (2.41 mmol, 419 µL, 1.2 eq.) and phenylpropionyl chloride (2 mmol, 307 µL, 1 eq.) were added under inert atmosphere, and the solution was stirred for 5 h at 23 °C. The mixture was concentrated using a rotary evaporator and the crude material was purified using a CombiFlash NextGen 300+ with reversed phase column C18 15.5 g Gold at a flow rate 13 mL per min with $H_2O$ (Solvent A) and ACN (solvent B) with the following gradient: 0-5 min, 5% B; 5-14 min, 40% B; 14-20 min 40% B; 20-25 min, 80% B. 5-ASA-phenylpropionic acid eluted at 15 min, 40% B. $^1$H-NMR (MeOD) δ 2.61 (t, 3H), 9.98 (t, 3H), 6.76 (2, 1H), 7.11-7.30 (m, 5H), 7.41 (d, 1H), 7.99 (d, 1H) ($^1$H-NMR spectra is available 10.5281/zenodo.17519052).

**Ibuprofen-carnitine**

Solid ibuprofen (4.85 mmol, 1 g, 1 eq.) and 3 mL of THF were added to a 20 mL vial with a stir bar and the reaction was placed in an ice/water bath, neat ethyl-chloroformate (5.82 mmol, 559 μL, 1.2 eq.) and triethylamine (7.27 mmol, 1.01ml, 1.5 eq.) were subsequently added, and the solution was stirred for 1.5 h, solid carnitine (4.85 mmol, 958 mg, 1 eq.) dissolved in 2 mL THF was subsequently added to the ibuprofen mixture. The reaction was removed from the ice/water bath and stirred overnight at 23 °C. The mixture was concentrated en vaccuo and purified using a CombiFlash NextGen 300+ with reversed phase column C18 15.5 g Gold at a flow rate13 mL per min with $H_2O$ (Solvent A) and ACN (solvent B) using the gradient: 0-2 min, 5% B; 3-10 min, 20-40% B; 11-14 min 60% B; 15-17 min, 80% B. Ibuprofen-carnitine eluted at 3-10 min, 20% B. . $^1$H-NMR (CDCl3) δ 0.86 (m, 6H), 1.73 (m, 3H), 1.18 (m, 1H), 2.22 (m, 2H), 2.42 (dd, 2H), 2.85-3.11 (m, 9H), 3.42 (m, 1H), 5.5 (m, 1H), 7.08-7.14 (m, 4H) ($^1$H-NMR spectra is available 10.5281/zenodo.17519052).

**Quantification**

Quantification of Ibuprofen-carnitine and 5-ASA-phenylpropionic acid was performed from urine sample and fecal sample, respectively (scripts used for quantification are available (abubakerpatan/Quantification-Script: Script used for quantification).

The LC-MS/MS method used for the analyses of the method validation and quantification was the same as previously described in LC-MS/MS data collection. The analytical method was performed according to the International Conference on Harmonization (ICH) guidelines[41] for ibuprofen carnitine and 5-ASA phenylpropionic acid. The method was validated based on the evaluation of the following parameters: specificity, precision (repeatability and intermediate precision), linearity, limit of detection (LOD), limit of quantification (LOQ), and accuracy. A matrix match calibration curve was created by spiking pool urine (ibuprofen carnitine) and pool fecal (5-ASA phenylpropionic acid) into calibrates to create a matrix match calibration curve for quantitation. Detailed information regarding the methodology used for each of them is described below. The validation was performed using Rise Plus Urobiome samples of human urine MSV000096359 that would contain the ibuprofen carnitine compound and Crohn's cohort MSV000099375 that contains the 5-ASA phenylpropionic acid. Peak area for ibuprofen carnitine and 5-ASA phenylpropionic acid was extracted using Skyline[42] (version 23.1). The method employed reached the acceptance criteria specified for each parameter of ibuprofen carnitine (Table 1), and 5-ASA phenylpropionic acid (Table 2). For quantification in biological samples, one sample of the Crohn's cohort and one sample of the Rise Plus Urobiome study with the highest peak area was injected in the validated method (samples were resuspended in 100 μL of 50/50 MeOH/$H_2O$ containing 1 μM of sulfamethazine). For the calculation of the amounts in the samples, it was estimated that 200 uL of urine sample and 54 mg fecal samples would be the starting material, and the extraction yield was also extrapolated to 100%.

The specificity was determined by injecting a blank solution containing only the internal standard (sulfadimethazine), and an injection of a solution containing all the ibuprofen carnitine (n=3). The relative standard deviation (RSD) was calculated based on each peak's retention time in the Rise Plus Urobiome and fecal samples. The MS and MS/MS spectra confirmed the specificity and identity of these compounds. The retention times of the peak of interest were as follows: Ibuprofen carnitine, 2.09 min and 5-ASA-phenylpropionic acid 4.52 min. These compounds didn't show interferences compared to the solution containing only the mixture of standards.

The linearity of the method was determined by calibration curves in concentration ranges comprising each compound at the samples of interest. A stock solution containing 1mM of each Ibuprofen carnitine and 5-ASA phenylpropionic acid was prepared in 50/50 MeOH/$H_2O$, followed by serial dilutions to get the concentration range mentioned in (Table 1) and (Table 2) and used to acquire calibration curves for all the compounds simultaneously. From this solution, 7 points were prepared with levels ranging from 10nM to 1uM for Ibuprofen carnitine and 100nM to 2uM for 5-ASA phenylpropionic acid with each spike with urine matrix Ibuprofen carnitine and fecal matrix for 5-ASA phenylpropionic acid. Each concentration level was injected in triplicate and the analytical curves were built based on the nominal concentrations, and the average between the ratios of each compound and the internal standard used (Ratio = $A_{compound}/A_{IS}$). A polynomial equation was obtained for each curve, and the correlation coefficients (R) were calculated for each compound. The R coefficients are available in (Table 1) and (Table 2).

LODs and LOQs were estimated by the mean of the slopes (S) and the standard deviation of the y-intercept (y). These limits were calculated by the following equations: LOD = (3.3*y)/S and LOQ = (10*y)/S. All the slopes, intercepts, LODs, and LOQs are shown in (Table 1) and (Table 2). The accuracy and precision of the method was determined by recovery analyses. For this, known amounts of the solution containing the standards were spiked to the sample P1-D-6_2_5753 for 5-ASA phenylpropionic acid and sample STD_SPK_urine sample for Ibuprofen carnitine solutions in two different concentrations (low and high) considering the predetermined calibration curve and concentration range. Three replicates for each level were injected and analyzed in the validated method. The accuracy was determined by the difference between the theoretical and experimental concentration values and the values were within the acceptance range of 80–120% and the precision by coefficient variation (CV).

**Data availability**

All data in this study are publicly available and accessible on GNPS/MassIVE. The multiplex synthesis library is available at https://external.gnps2.org/gnpslibrary ("MULTIPLEX-SYNTHESIS-LIBRARY", 6 partitions in total). All untargeted metabolomics LC-MS/MS data have been deposited to GNPS/MassIVE under the accession numbers MSV000097885, MSV000097874, MSV000097869, MSV000094559, MSV000094447, MSV000094393, MSV000094391, MSV000094382, MSV000094337, MSV000094300, MSV000098637 (bile acid), MSV000098628 (small molecules), MSV000098639 (drug compounds), MSV000098640 (peptides), MSV000096359 (Rise Plus Urobiome samples), and MSV000099150 (urine from 9 pregnant women), MSV000099374 (data files for standards and biological samples for retention time matching of 5-ASA, ibuprofen and succinic acid conjugates), MSV000099375 (5-ASA quantification data), MSV000099556 (ibuprofen quantification data). The job link for searching the multiplex synthesis library against existing GNPS libraries is available at https://gnps2.org/status?task=0e77aa138fc2473ab8a801a8d59905e6. The classical molecular network of synthetic MS/MS spectra that are exclusively present in humans is available at https://gnps2.org/status?task=a6b9129f880146b0aef3168855c32713. The FBMN jobs of three datasets used for 5-ASA-phenylpropionic acid can be accessed using the following links: IBD dataset (MSV000082094): https://gnps2.org/status?task=5f230f976ccb4f19aa94d59407468138; Pediatric IBD cohort (MSV000097610): https://gnps2.org/status?task=023988a1842146d6a5b2ba87a3212598; Rheumatoid arthritis cohort (MSV000084556): https://gnps2.org/status?task=16da6e571d574a829e3de75dd610bc97.

## Code availability

Source codes for all the data analyses applied on the multiplex synthesis library can be found at https://github.com/Philipbear/multiplex_synthesis. The tool for generating SMILES strings of multiplex synthesis products can be accessed at https://autosmiles.streamlit.app/rxnSMILES. All scripts used for quantification are available at https://github.com/abubakerpatan/Quantification-Script. The codebase of the MS/MS library generation workflow is available at https://github.com/Wang-Bioinformatics-Lab/Reverse_metabolomics_library_generation. All scripts used to generate the figures in this study can be accessed from GitHub (https://github.com/Philipbear/multiplex_synthesis and https://github.com/VCLamoureux/synthesis_multiplex).

## Acknowledgements

PCD acknowledges NIDDK R01DK136117, U24DK133658, and EnvedaGives Scientific Research Fund for enabling this work. MW acknowledges NIGMS U24DK133658. AP was supported by R01DK136117. SX was supported by BBSRC/NSF award 2152526. VCL was supported by Fonds de recherche du Québec - Santé (FRQS) postdoctoral fellowship (335368) and from Natural Sciences and Engineering Research Council of Canada (NSERC) postdoctoral fellowship (598938). NB and JC were partially supported by NIDDK U24DK133658 and NIGMS R24GM148372. HM-R is supported by U24DK133658. LCD is supported by the DDRA (Doctoral Dissertation Research Award) from the Fulbright U.S. Student Program, which is sponsored by the U.S. Department of State and the Fulbright Brazil Commission. AMC-R and PCD were supported by the Gordon and Betty Moore Foundation grant GBMF12120. Reproductive Scientist Development Program Grant supports LAB ACTRI Pilot Project Grant to LAB.

## References

1. El Abiead, Y. *et al.* Discovery of metabolites prevails amid in-source fragmentation. *Nat. Metab.* **7**, 435–437 (2025).
2. El Abiead, Y. *et al.* A Perspective on Unintentional Fragments and their Impact on the Dark Metabolome, Untargeted Profiling, Molecular Networking, Public Data, and Repository Scale Analysis. Preprint at https://doi.org/10.26434/chemrxiv-2025-l6pg7 (2025).
3. Xing, S. *et al.* Reverse Spectral Search Reimagined: A Simple but Overlooked Solution for Chimeric Spectral Annotation. *Anal. Chem.* **97**, 17926–17930 (2025).
4. Charron-Lamoureux, V. *et al.* A guide to reverse metabolomics—a framework for big data discovery strategy. *Nat. Protoc.* 1–34 (2025) doi:10.1038/s41596-024-01136-2.
5. 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.
6. 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).
7. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
8. Batsoyol, N., Pullman, B., Wang, M., Bandeira, N. & Swanson, S. P-massive: a real-time search engine for a multi-terabyte mass spectrometry database. in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* 1–15 (IEEE Press, Dallas, Texas, 2022).
9. Yurekten, O. *et al.* MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* **52**, D640–D646 (2024).
10. 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).
11. El Abiead, Y. *et al.* Enabling pan-repository reanalysis for big data science of public metabolomics data. *Nat. Commun.* **16**, 4838 (2025).
12. 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.
13. West, K. A., Schmid, R., Gauglitz, J. M., Wang, M. & Dorrestein, P. C. foodMASST a mass spectrometry search tool for foods and beverages. *Npj Sci. Food* **6**, 22 (2022).
14. Gomes, P. W. P. *et al.* plantMASST - Community-driven chemotaxonomic digitization of plants. 2024.05.13.593988 Preprint at https://doi.org/10.1101/2024.05.13.593988 (2024).
15. Zuffa, S. *et al.* A Multi-Organ Murine Metabolomics Atlas Reveals Molecular Dysregulations in Alzheimer's Disease. 2025.04.28.651123 Preprint at https://doi.org/10.1101/2025.04.28.651123 (2025).
16. Bittremieux, W. *et al.* Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. 2020.05.09.086066 Preprint at https://doi.org/10.1101/2020.05.09.086066 (2020).
17. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
18. Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
19. Schymanski, E. L. *et al.* Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J. Cheminformatics* **13**, 19 (2021).

20. FooDB. https://foodb.ca/.

21. Scheubert, K. *et al.* Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).

22. Kim, H. W. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **84**, 2795–2807 (2021).

23. Mehta, R. S. *et al.* Gut microbial metabolism of 5-ASA diminishes its clinical efficacy in inflammatory bowel disease. *Nat. Med.* **29**, 700–709 (2023).

24. Crouwel, F., Buiter, H. J. C. & de Boer, N. K. Gut Microbiota-driven Drug Metabolism in Inflammatory Bowel Disease. *J. Crohns Colitis* **15**, 307–315 (2021).

25. Pruss, K. M. *et al.* Host-microbe co-metabolism via MCAD generates circulating metabolites including hippuric acid. *Nat. Commun.* **14**, 512 (2023).

26. Haffner Jacob J. *et al.* Untargeted Fecal Metabolomic Analyses across an Industrialization Gradient Reveal Shared Metabolites and Impact of Industrialization on Fecal Microbiome-Metabolome Interactions. *mSystems* **7**, e00710-22 (2022).

27. Murphy, M. P. & O'Neill, L. A. J. Krebs Cycle Reimagined: The Emerging Roles of Succinate and Itaconate as Signal Transducers. *Cell* **174**, 780–784 (2018).

28. Weinert, B. T. *et al.* Lysine Succinylation Is a Frequently Occurring Modification in Prokaryotes and Eukaryotes and Extensively Overlaps with Acetylation. *Cell Rep.* **4**, 842–851 (2013).

29. Wei, Y., Ma, X., Zhao, J., Wang, X. & Gao, C. Succinate metabolism and its regulation of host-microbe interactions. *Gut Microbes* **15**, 2190300 (2023).

30. Shirley, M. A., Guan, X., Kaiser, D. G., Halstead, G. W. & Baillie, T. A. Taurine conjugation of ibuprofen in humans and in rat liver in vitro. Relationship to metabolic chiral inversion. *J. Pharmacol. Exp. Ther.* **269**, 1166–1175 (1994).

31. Mohammed, H. O., Almási, A., Molnár, S. & Perjési, P. The Intestinal and Biliary Metabolites of Ibuprofen in the Rat with Experimental Hyperglycemia. *Molecules* **27**, (2022).

32. Castillo, M., Lam, Y. W. F., Dooley, M. A., Stahl, E. & Smith, P. C. Disposition and covalent binding of ibuprofen and its acyl glucuronide in the elderly. *Clin. Pharmacol. Ther.* **57**, 636–644 (1995).

33. Mazaleuskaya, L. L. *et al.* PharmGKB summary: ibuprofen pathways. *Pharmacogenet. Genomics* **25**, (2015).

34. Wang, G. *et al.* Intestinal OCTN2- and MCT1-targeted drug delivery to improve oral bioavailability. *Emerg. Role Transp. Drug Interact. Deliv.* **15**, 158–172 (2020).

35. Diao, L. & Polli, J. E. Synthesis and in vitro characterization of drug conjugates of l-carnitine as potential prodrugs that target human Octn2. *J. Pharm. Sci.* **100**, 3802–3816 (2011).

36. Wang, R. *et al.* Neural Spectral Prediction for Structure Elucidation with Tandem Mass Spectrometry. *bioRxiv* 2025.05.28.656653 (2025) doi:10.1101/2025.05.28.656653.

37. Shahneh, M. R. Z. *et al.* ModiFinder: Tandem Mass Spectral Alignment Enables Structural Modification Site Localization. *J. Am. Soc. Mass Spectrom.* https://doi.org/10.1021/jasms.4c00061 (2024) doi:10.1021/jasms.4c00061.

38. Bittremieux, W. *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nat. Commun.* **14**, 8488 (2023).

39. Weldon, K. C. *et al.* Urinary Metabolomic Profile is Minimally Impacted by Common Storage Conditions and Additives. *Int. Urogynecology J.* **36**, 839–847 (2025).

40. Mongia, M. *et al.* Fast mass spectrometry search and clustering of untargeted metabolomics data. *Nat. Biotechnol.* **42**, 1672–1677 (2024).

41. ICH Q10. International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use. (2005).

42. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).