Empirically establishing drug exposure records directly from untargeted metabolomics data

Haoqi Nina Zhao[1,2,†], Kine Eide Kvitne[2,3,†], Corinna Brungs[4†], Siddharth Mohan[2], Vincent Charron-Lamoureux[1,2], Wout Bittremieux[1,2,5], Runbang Tang[2], Robin Schmid[1,2,4], Santosh Lamichhane[2,6], Yasin El Abiead[1,2], Mohammadsobhan S. Andalibi[7,8,9], Helena Mannochio-Russo[1,2], Madison Ambre[10], Nicole E. Avalon[11], MacKenzie Bryant[10], Andrés Mauricio Caraballo-Rodríguez[1,2], Martin Casas Maya[10], Loryn Chin[12], Ronald J. Ellis[7,8], Donald Franklin[8], Sagan Girod[13], Paulo Wender P Gomes[1,2,14], Lauren Hansen[10], Robert Heaton[8], Jennifer E. Iudicello[8], Alan K. Jarmusch[1,2,15], Lora Khatib[7], Scott Letendre[9,16], Sarolt Magyari[2,17], Daniel McDonald[10], Ipsita Mohanty[1,2], Andrés Cumsille[2,18], David J. Moore[8,9], Prajit Rajkumar[2], Dylan H. Ross[19,20], Harshada Sapre[2], Mohammad Reza Zare Shahneh[21], Sydney P. Thomas[1,2], Caitlin Tribelhorn[10], Helena M. Tubb[10], Corinn Walker[10], Crystal X. Wang[8,9], Shipei Xing[1,2], Jasmine Zemlin,[1,2,22] Simone Zuffa[1,2], David S. Wishart[12,23], Rima Kaddurah-Daouk[24,25,26], Mingxun Wang[21], Manuela Raffatellu[10,22,27], Karsten Zengler[11,10,22,28], Tomáš Pluskal[4], Libin Xu[19], Rob Knight[10,22,29,30,31], Shirley M. Tsunoda[2], Pieter C. Dorrestein[1,2,22*]

[1] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[2] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[3] Department of Pharmacy, University of Oslo, Oslo, Norway
[4] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic
[5] Department of Computer Science, University of Antwerp, Antwerp, Belgium
[6] Turku Bioscience Centre, University of Turku and Åbo Akademi University, Tykistönkatu 6A, 20520 Turku, Finland
[7] Department of Neurosciences, University of California San Diego, La Jolla, CA, USA
[8] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA
[9] HIV Neurobehavioral Research Program, University of California San Diego, La Jolla, CA, USA
[10] Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
[11] Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA
[12] Department of Bioengineering, University of California San Diego, La Jolla, California, USA.
[13] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada
[14] Faculty of Chemistry, Federal University of Pará, Belém, PA, Brazil
[15] Immunity, Inflammation, and Disease Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA
[16] Department of Medicine, University of California San Diego, La Jolla, CA, USA.
[17] Institute of Microbiology, Eidgenössische Technische Hochschule (ETH) Zürich, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland
[18] Department of Microbiology and Cell Sciences, University of Florida, Museum Drive, Gainesville, FL, USA

[19] Department of Medicinal Chemistry, University of Washington, Seattle, WA, USA
[20] Current address: Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA
[21] Department of Computer Science and Engineering, University of California Riverside, Riverside, CA, USA
[22] Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA.
[23] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada
[24] Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, 27708, USA
[25] Duke Institute of Brain Sciences, Duke University, Durham, NC, USA
[26] Department of Medicine, Duke University, Durham, NC, USA
[27] Chiba University, UC San Diego Center for Mucosal Immunology, Allergy, and Vaccines (CU-UCSD cMAV), La Jolla, CA, USA
[28] Program in Materials Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0418, USA
[29] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[30] Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[31] Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
[†] Haoqi Nina Zhao, Kine Eide Kvitne, and Corinna Brungs contributed equally to this work.
* Author to whom correspondence should be addressed.

Abstract

Despite extensive efforts, extracting information on medication exposure from clinical records remains challenging. To complement this approach, we developed the tandem mass spectrometry (MS/MS) based GNPS Drug Library. This resource integrates MS/MS data for drugs and their metabolites/analogs with controlled vocabularies on exposure sources, pharmacologic classes, therapeutic indications, and mechanisms of action. It enables direct analysis of drug exposure and metabolism from untargeted metabolomics data independent of clinical records. Our library facilitates stratification of individuals in clinical studies based on the empirically detected medications, exemplified by drug-dependent microbiota-derived *N*-acyl lipid changes in a human immunodeficiency virus cohort. The GNPS Drug Library holds potential for broader applications in drug discovery and precision medicine.

Main Text

Growing evidence suggests that the chemical exposome plays a critical role in shaping human health, with drugs being a significant source of chemical exposure that carries profound health implications.[1] According to a recent survey by the Center for Disease Control and Prevention, nearly half (45.7%) of the U.S. population used at least one prescription drug in the past 30 days.[2] Drug concentrations in human blood are on par with those of endogenous and dietary molecules,[3] and they have important impacts on the metabolic states and microbiome composition.[4–7] Clinical research typically relies on medical records or self-reporting surveys to assess drug exposure,[8] but these methods are costly and often incomplete.[8–10] They often overlook over-the-counter medications and supplements, and fail to account for patient adherence. Additionally, they miss drug usage not documented in medical records, such as those purchased online,[10,11] acquired across borders,[12,13] or consumed through secondary use of leftover drugs. Medical records are also incapable of documenting drugs introduced into the food supply that are unknowingly consumed, such as the antifungal natamycin used both to treat fungal eye infections and as a preservative for dairy products. Additionally, the varying half-lives of drugs and their metabolites further complicate exposure assessment, as some drugs are rapidly eliminated from the body while others can persist for months.[14,15]

Untargeted metabolomics offers the opportunity to complement clinical records by empirically establishing the presence of drugs and their metabolites directly from biological samples. However, liquid chromatography-tandem mass spectrometry (LC-MS/MS) based annotations, which rely on reference MS/MS library matches, are difficult to interpret. For example, annotation may return a complex IUPAC chemical name like "(2R,3S,4R,5R,8R,10R,11R,12S,13S,14R)-11-[(2S,3R,4S,6R)-4-(dimethylamino)-3-hydroxy-6-methyloxan-2-yl]oxy-2-ethyl-3,4,10-trihydroxy-13-[(2R,4R,5S,6S)-5-hydroxy-4-methoxy-4,6-dimethyloxan-2-yl]oxy-3,5,6,8,10,12,14-heptamethyl-1-oxa-6-azacyclopentadecan-15-one". A text search in the right reference resource or an open web search can hopefully link this IUPAC name to "azithromycin", the name used in clinical settings. A second search of the term "azithromycin" is then required to connect the name to its function as a drug, in this case an antibiotic originally isolated from a bacterium. While this example involves a simple name and a limited number of identifiers for azithromycin, other compounds, like penicillin G or aspirin, have hundreds of synonyms and identifiers in chemistry databases such as PubChem, making the identification process more challenging. This task must be repeated for every obtained annotation, which can range from hundreds to thousands in a given metabolomics experiment, to find all the drugs in a dataset.

Even when the drugs are identified, the interpretation of their biological impacts requires extensive literature and web searches to understand the therapeutic indications of the drugs and their mechanisms of action. Public databases, such as DrugBank,[16,17] DrugCentral,[18] DailyMed,[19] and KEGG DRUG,[20] can assist in interpretation, but the pharmacologic information is often provided as plain text or combinatorial classifications that require manual organization before downstream analysis. Although, in principle, large language models or similar text mining strategies can assist in this, the results of such models still need manual verification to confirm accuracy.[21–23] In addition, it is not uncommon that only metabolized versions of a drug is present in the sample, leading to missed drug exposure if only the parent drug is considered.[24] Unfortunately, with very few exceptions, reference MS/MS libraries include only the parent drug

but not the drug metabolites due to challenges in obtaining reference standards for these metabolites.

The absence of MS/MS spectra for many drug metabolites, along with other challenges mentioned above, makes it very difficult to accurately annotate clinical samples for all drug exposures. For instance, stratifying a cohort based on antibiotic exposure - perhaps to better understand microbiome changes or as an exclusion criterion for clinical studies - requires identifying all antibiotics and their metabolites present in the samples. This is currently challenging due to the lack of resources that provide objective, systematic, and efficient readouts of drugs in untargeted metabolomics experiments.

To address this gap and to enable data science strategies on drug readouts, we curated the Global Natural Product Social Molecular Networking (GNPS) Drug Library, a collection of reference spectra for drugs and their metabolites/analogs (including parent ion masses and MS/MS spectra) along with structured pharmacologic metadata including exposure source, pharmacologic class, therapeutic indication, and mechanism of action. This comprehensive resource will enable further data science analysis to empirically - and retroactively - determine drug exposure using untargeted metabolomics data, complementing the information available in clinical records.

The creation of this library involved three key steps: 1) collecting MS/MS spectra of drugs and drug metabolites from publicly available MS/MS reference libraries; 2) finding MS/MS spectra analogs of those drugs in publicly accessible untargeted metabolomics data to enhance coverage of the metabolized versions of drugs; and 3) linking each MS/MS spectrum of a drug to controlled-vocabulary metadata - the key component of this resource that facilitates efficient data interpretation (**Figure 1a**).

The reference MS/MS spectra of drugs and their known metabolites were collected from two of the largest open-access mass spectral libraries, namely the GNPS Spectral Library[25] and MS^nLib[26]. For all the MS/MS spectra in the GNPS and MS^nLib, metadata enrichment was first performed against PubChem (for synonyms),[27] ChEMBL, DrugCentral,[18] the Broad Institute Drug Repurposing Hub databases[28] (for pharmacological information),[29] and DrugBank (for pharmacological information and the Anatomical Therapeutic Chemical Classification code).[16,17,30] This process utilized the available metadata in the GNPS Spectral Library and MS^nLib, including the chemical structures (e.g., SMILES or InChI), database identifiers (e.g., DrugBank ID or ChEMBL ID), and compound names. Based on the enriched metadata regarding clinical phases, all MS/MS spectra of drugs and compounds in clinical trials were compiled into the centralized GNPS Drug Library (see method details in **Supplementary Text 1**), resulting in 99,122 MS/MS reference spectra for 4,723 unique compounds. The compound names in the GNPS Drug Library were automatically curated and set to the first synonym in PubChem. We note that the term “drug” is used here in a broad sense, as the GNPS Drug Library includes not only prescribed and over-the-counter medications but also compounds currently in clinical trials, drugs that have been withdrawn, as well as substances with potential for abuse (e.g., cocaine, fentanyl).

Given that drug metabolites are largely overlooked in the initial search, we performed a second “partial name match” to include metabolites that retain the full drug names. For example, by searching for the name “venlafaxine” (a serotonin and norepinephrine reuptake inhibitor used to treat depression and various anxiety disorders), we obtained reference spectra for five of its metabolites, including “*N*-desmethylvenlafaxine”, “*O*-desmethylvenlafaxine”, “*N*,*O*-

didesmethylvenlafaxine", "*N*,*N*-didesmethylvenlafaxine", and "venlafaxine *N*-oxide". Using this strategy combined with manual inspection of the results, we captured 2,080 reference spectra for the metabolites of 110 drugs. Lastly, we added the MS/MS spectra collected in the development of dmCCS,[31] a collision cross section database for drugs and their metabolites where human liver microsomes and S9 fraction were used for *in vitro* generation of drug metabolites. In total, 4,087 spectra for the metabolites of 470 drugs were included in the GNPS Drug Library (**Figure 1a**).

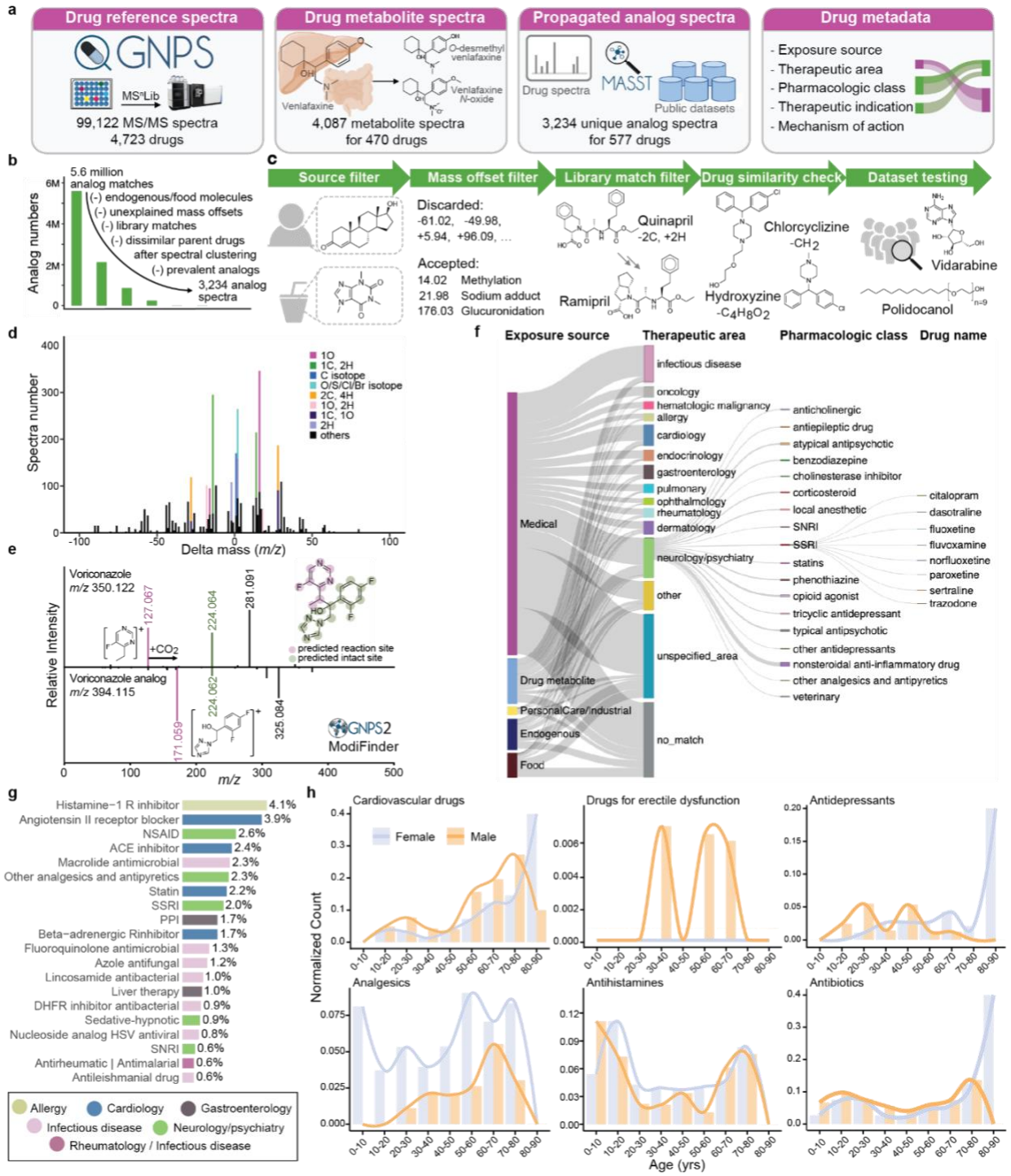

Figure 1. The GNPS Drug Library and connected pharmacologic metadata. a, The GNPS Drug Library comprises four key resources: Drug MS/MS reference spectra, drug metabolite MS/MS reference spectra, propagated drug analogs derived from public metabolomics datasets, and pharmacologic metadata connected to each reference spectrum. **b,** FastMASST analog search of drug spectra against public metabolomics studies yielded propagated drug-analogous MS/MS spectra, which were filtered by removing analogs for drugs with endogeneous and food sources (source filter), removing mass offsets unexplained by of common metabolic pathways (mass offset filter), removing analogs with GNPS library matches (library match filter), removing analogs connected to multiple drugs with dissimilar structures after spectra clustering (drug similarity filter), and removing analogs with unrealistic drug exposure indications (dataset testing). **c**, Illustration of each filter employed in curating FastMASST analog match results. **d**, Frequency of mass offsets in the propagated drug analog library. The mass offsets were grouped by unit mass and stacked based on the number of analog spectra. The most frequently observed mass offsets are colored while the rests are black. **e,** An example of structural modification sites predicted by ModiFinder.[32] Purple color highlights modified spectra and substructures, while the green color highlights unmodified ones. **f**, Overview of the ontology-based drug metadata, highlighting common pharmaceutical classes and specific drugs in the neurology/psychiatry category. Width of the bars and lines reflects the number of unique drug structures in each class. **g,** The top 20 most detected pharmacologic classes in fecal samples from the American Gut Project,[33] a cohort of the general population from the United States (US), Europe, and Australia (1,993 individuals). **h,** Detected therapeutic drug class patterns by age and sex (1,845 individuals with age and sex information; age 46 ± 18 years [range 3-93], with 53% being female). Detection of cardiovascular drugs increased with age, while analgesics, antihistamines, and antibiotics were detected across all ages.[34,35] Analgesics were more frequently detected in females, consistent with the literature,[36,37] and drugs for erectile dysfunction were detected only in males. NSAID, non-steroidal anti-inflammatory drugs; ACE, angiotensin converting enzyme; SSRI, selective serotonin reuptake inhibitor; PPI, proton pump inhibitor; DHFR, dihydrofolate reductase; HSV, herpes simplex virus; SNRI, serotonin and norepinephrine reuptake inhibitor.

Despite the extensive collection effort, metabolite reference spectra were only available for 10% of the drugs included in the GNPS Drug Library. We hypothesized that unannotated drug metabolites are present in public untargeted metabolomics data. We further hypothesized that spectral alignment strategies can be used to find the modified versions of the drugs.[38–40] In another word, public untargeted metabolomics data could be used to create a reference library of candidate drug metabolites that will facilitate the drug exposure readout in future datasets.

Based on MS/MS spectral alignment using two computational methods: repository-scale molecular networking[41] and fastMASST with analog search,[42,43] we retrieved all possible MS/MS spectra analogous to drugs from the GNPS/MassIVE public repository (~2,700 LC-MS/MS datasets).[25] These spectra represent drug-related molecules potentially derived from metabolism (host or microbiome), abiotic processes, and adducts of drugs from MS measurements. We obtained analogous MS/MS spectra for 14.6% of the 103,209 reference spectra for drug and drug metabolites (>5.5 million drug-analog spectral pairs).

In testing of the propagated analog library, we identified the need for additional filters to enhance its relevance to drug exposure (**Figure 1b-c**). First, it is not possible to determine the sources of exogenously supplied chemicals that are also produced endogenously or derived from the diet. Consequently, structural analogs of drugs with endogenous or dietary sources were excluded from the propagated drug analog library (e.g., analogous MS/MS spectra of testosterone used to treat hypogonadism, or caffeine used as a stimulant drug, were excluded). Second, propagated analogs with uncommon or unexplained mass offsets (precursor mass difference

between the propagated analog and the connected drug) were excluded. Mass offsets were obtained from UNIMOD,[44] from a community-curated list of explainable delta masses (**Table S1**), and from the Host Gut Microbiota Metabolism Xenobiotics Database,[45] and were manually curated for those relevant to drug metabolism (e.g., 14.02 Da, methylation; 176.03 Da, glucuronidation) or mass spectrometry adducts (e.g., 17.03 Da, ammonium adduct; see **Table S2** for the 156 mass offsets that were included). MS/MS spectra of propagated analogs were excluded if the mass offsets were not in the customized list, or when the mass offsets occurred fewer than ten times. Third, since drugs within the same pharmacologic family often have similar structures, they can be identified as analogs of each other through spectral alignments. Therefore, we excluded MS/MS spectra with matches to the GNPS library from the propagated analog annotations. For example, a propagated analog of quinapril, an angiotensin converting enzyme (ACE) inhibitor, had a spectral match to ramipril, another ACE inhibitor (**Figure 1c**). Excluding these analog annotations ensures that they do not overwrite library matches of known drugs and metabolites. Fourth, if one propagated analog spectrum is connected to multiple drugs after spectral clustering, the drugs need to be structurally similar to accept the shared analog. We illustrate this with the propagated analog (*m/z* 287.133, with a formula $C_{17}H_{19}ClN_2$) that is connected to both hydroxyzine [$C_{21}H_{27}ClN_2O_2$, mass offset 88.05 Da ($C_4H_8O_2$)] and chlorcyclizine [$C_{18}H_{21}ClN_2$, mass offset 14.02 Da (CH_2)], which share the core structure (**Figure 1c**). Finally, we tested the propagated drug analog library against 10 public LC-MS/MS datasets to filter out analogs that have unrealistic drug exposure indications. The selected datasets represent a broad range of human tissue types and biofluids, including four fecal, two breast milk, one plasma, and one brain dataset, as well as two datasets containing multiple mouse tissues with metadata confirming no drugs were used. We observed analogs of tocofersolan (a synthetic vitamin E derivative), iloprost (a synthetic prostacyclin mimetic), desonide (a synthetic topical corticosteroid), medroxyprogesterone (a synthetic progestin), and vidarabine (an adenosine analog used as an antiviral) in >50% of the human fecal samples from the American Gut Project (n = 1,993 data files), a cohort of the general population. The connected drugs for these analogs are derivatives of endogenous or food derived molecules, and are unlikely to be used by more than half of the population. Therefore, these analogs cannot be confidently linked to drug exposures and were excluded. Analogs of polidocanol (a synthetic long-chain fatty alcohol used as anesthetics) were observed in >70% of 2,463 human milk samples. They are likely surfactants/contaminants with the polyethylene glycol structural units[46] and thus were excluded from the propagated drug analog library (**Figure 1c**).

After all filtering steps, 3,234 clustered MS/MS spectra representing propagated analogs of 577 drugs were retained in the final drug analog library. We observed that 75% of the propagated analogs occurred at least once in the same data file with the corresponding parent drugs (**Table S3**). The most common mass offsets in the drug analog library correspond to a gain of 15.99 Da, which can be interpreted as the gain of an oxygen (e.g., oxidative metabolism of the drug), followed by a gain or loss of 14.02 Da (CH_2, (de)methylation), a gain of 1.00 Da (13C isotope), a gain of 2.00 Da (O/S/Cl/Br/2C isotopes), and a gain or loss of 28.03 Da (C_2H_4, (de)ethylation; **Figure 1d**). Notably, it is possible that such analog spectra are MS/MS of other ion forms of the parent drug, such as isotopes, different adducts, in/post-source fragments, or multimers, rather than drug metabolites or structural analogs. However, their indications in drug exposure remain the same and thus we did not separate drug metabolites and instrument adducts

in drug exposure stratification. To extend structural hypotheses for the drug analogs, we employed the newly developed ModiFinder,[32] which leverages the shifted MS/MS fragment peaks in the MS/MS alignment to predict the most likely location for the structural modifications. We were able to predict the partial location of the modification for 61.5% of the analog spectra. We demonstrate examples where ModiFinder predictions agree with expert manual interpretation of the MS/MS spectra (**Figure 1e, S1**).

Connecting drug detections to their therapeutic indications typically requires expert knowledge and/or extensive literature searches. The GNPS Drug Library addresses this challenge by providing controlled-vocabulary metadata together with the specific drug annotations. This allows users to annotate all drugs in an untargeted metabolomics dataset and directly obtain a table with exposure sources, pharmacologic classes, therapeutic indications, and mechanisms of action of the drugs, with their structures and names in a data science ready format (**Figure 1f, S2**). Particularly, the "exposure source" information categorizes the drugs in a combination of five classes, namely medical, endogenous, food, personal care, and industrial sources, which was based on the source categorizations from the Chemical Functional Ontology (ChemFOnt) database[47] and modified manually - by parsing of web pages and scientific literature - to increase compound coverage and improve accuracy and consistency. This categorization allows distinguishing endogenous or food sourced molecules (for the non-analogous spectra only). Examples include deoxycholic acid, an endogenous molecule also used for liver disease, and lactitol, a food sweetener also used as a laxative. Using the GNPS Drug Library metadata, such annotations can be separated from those molecules used exclusively as drugs, which have entirely different exposure implications.

Through structural and name matches, we extracted the pharmacologic classes of 900 drugs from the U.S. Food and Drug Administration (FDA) and the therapeutic areas, therapeutic indications, and mechanisms of action for 3,894 drugs from the Broad Institute Drug Repurposing Hub.[28] However, we noticed substantial variability in the extracted information (e.g., inconsistent therapeutic areas assigned to drugs within the same pharmacological class; the sulfonamide antimicrobials sulfamethizole, sulfamethazine, and sulfacetamide were categorized as infectious disease, gastroenterology, and ophthalmology, respectively), or insufficient metadata for several drugs (e.g., common therapeutic indications missing). Therefore, this metadata was further manually curated by expert clinical pharmacologists to enhance and clean up the information retrieved from databases. This manual curation increased the metadata coverage to 4,560 drugs. Drugs without associated metadata are typically those that have been withdrawn from the market (e.g., indoprofen), were in drug development but never marketed (e.g., tarafenacin), or are under development but do not yet have regulatory approval (e.g., firsocostat).

In total, 735 drugs in the GNPS Drug Library (38,001 spectra) were identified with endogenous or dietary sources. The final metadata of the drug library covers 27 unique therapeutic areas, 571 pharmacological classes, 920 therapeutic indications, and 823 mechanisms of action (**Figure 1f, S2**). Therapeutic areas of neurology/psychiatry, infectious disease, and cardiology have the highest number of included drugs (**Figure 1f**) and reference spectra (**Figure S2**). We note that these incidences reflect the availability of the reference spectra but not necessarily the prevalence of these drugs in the general population. Combining the exposure source and therapeutic area, we noticed that fewer drugs related to infection and neurology/psychiatry have endogenous or food sources, while higher portions of drugs used for

gastroenterology (e.g., deoxycholic acid, riboflavin) and dermatology (e.g., salicylic acid, nicotinamide) are endogenous and/or can come from food-derived molecules.

Connected with public untargeted metabolomics data, the GNPS Drug Library can reveal distinct drug exposure profiles among different disease, age, and gender groups. For different disease studies, we used the human disease ontology identifier (DOID) curated in ReDU, a controlled-vocabulary metadata for public metabolomics datasets,[48] and searched for the drugs and drug analogs using fastMASST.[42] Samples from individuals with inflammatory bowel disease, Kawasaki disease, and dental caries were characterized by high detection frequencies of antibiotics (**Figure S3a**). Skin swabs of patients with psoriasis were characterized by antifungals. Samples from people with human immunodeficiency virus (HIV) showed high frequency of antivirals, and samples from Alzheimer's disease patients were characterized by cardiology and neurology/psychiatry drugs, all consistent with the expected drug usage of persons in these disease groups (**Figure S3a**).

To investigate drug exposures among different age and sex groups, we profiled 1,993 fecal samples from the American Gut Project,[33] with participants from the United States (US), Europe, and Australia with age 46 ± 18 years (range 3-93; 53% female). A total of 75 different drugs were detected; the most frequently detected pharmacologic classes included histamine-1 receptor antagonist (allergy), angiotensin II-receptor blocker (cardiology), ACE inhibitor (cardiology), beta-adrenergic receptor inhibitor (cardiology), statin (lipid-lowering), non-steroidal anti-inflammatory drug (NSAID; analgesics), and selective serotonin reuptake inhibitor (SSRI; antidepressant), which matches with the most commonly prescribed drug classes in these regions (**Figure 1g**).[49–51] There were more drugs per individual noted in the US cohort compared to the European and Australian cohorts (chi-square test; χ^2 (8, n = 1,903) = 31, p = 0.00014, **Figure S3b**). When connected with age and sex information, the drug detection agrees with the expected usage patterns of different drug classes (**Figure 1h**). For example, cardiovascular drugs were detected more frequently with increasing age, while analgesics, antihistamines, and antibiotics were detected across all ages.[34,35] We also observed that analgesics, such as NSAIDs and paracetamol, were more frequently detected in females (chi-square test; χ^2 (1, n = 1,958) = 15.4, p = 8.54×10^{-5}), consistent with the literature,[36,37] and that drugs for erectile dysfunction were detected only in males. Overall, empirical drug readout using untargeted metabolomics, facilitated by the GNPS Drug Library, demonstrated good specificity among different disease, age, and sex groups.

The GNPS Drug Library can allow the discovery of previously uncharacterized drug metabolites. To illustrate this, we analyzed fecal samples from the HIV Neurobehavioral Research Center (HNRC) cohort (n = 322; average age 55 ± 12 years), which included both people with HIV (n = 222) and people without HIV (n = 100). Among the 17,729 unique MS/MS spectra obtained, 493 were annotated with the GNPS Drug Library. After removing drugs that could be from endogenous or food sources (because we cannot assess whether they were given as a medication) and grouping annotations of drugs, metabolites, or analogs, 169 unique drugs remained. Antiretroviral drugs (ARVs; drugs for the treatment of HIV), drugs for cardiovascular disease, and drugs for anxiety and depression are the most frequently detected categories (**Figure 2a, S4a**). Despite the high rates of viral suppression with the advent of antiretroviral therapies (ART; a combination of ARVs to treat HIV), people with HIV have disproportionately

high rates of depression and cardiovascular diseases,[52–55] reflected in the observation of antidepressants and cardiovascular drugs in these samples.

Interestingly, 33% of the drugs were annotated together with their metabolites or analogs, and the occurrences of drug metabolites/analogs aligned with those of the parent drugs (**Figure 2a**). For example, darunavir (an ARV) had no annotated metabolites but was observed with 10 analogs (**Figure S4b**). Retention time and peak shape analysis indicated that two of the darunavir analogs are in-source fragments (as judged by overlapping retention times),[56] while the others remain unknown metabolites or isomers of this drug (**Figure S4c,d**). For the analogs that are not in-source fragments, 63-100% (median 96%) of their occurrences were together with the darunavir parent drug. The observations of darunavir analogs without the parent drug are perhaps related to sample acquisition post conclusion of a dosing interval and metabolites staying in the body for prolonged periods of time, or to ultrarapid metabolism of the participants that impaired the detection of the parent drug. Nevertheless, this observation highlights the utility of drug metabolites and analogs to increase the sensitivity of drug exposure readouts via untargeted metabolomics. We note that the HNRC dataset was added to the GNPS/MassIVE public repository after the development of the drug analog library. Therefore, analog mining via existing public metabolomics datasets can facilitate the discovery of uncharacterized metabolites in new data.

To further investigate the potential metabolic sources of the observed drug analogs, we cultured darunavir and 12 other drugs with a defined and complex synthetic microbial community of 111 bacterial species commonly found in the human gut.[57] Except clindamycin (an antibiotic), all drugs observed with three or more metabolites/analogs and in >10% samples were incubated (10 drugs in total; **Table S4**); omeprazole, loratadine, and terbinafine were additionally included because their analogs were frequently observed in samples without parent drugs. Shared analogs were observed for 10 of the 13 drugs between the fecal samples and the microbial incubations. Among them, metabolites/transformation products were observed for 4 drugs (ritonavir, atorvastatin, abacavir, and omeprazole; **Figure 2b, S5**), while the rest of the analogs were in-source fragments based on retention time correlation analysis.[56] The ritonavir, atorvastatin, and abacavir analogs increased in intensity with increased microbial incubation time (**Figure S5a-d**), indicating microbial metabolism as a possible source and consistent with their observation in fecal samples. The omeprazole analog (*m/z* 330.127) appeared to be an abiotic transformation product because it was already present at t=0 cultures, and its intensity decreased with increased incubation time (**Figure S5e-g**). This is consistent with the fast activation of omeprazole (*m/z* 346.122), a proton-pump inhibitor and a prodrug, to the reactive sulphenamide product (*m/z* 330.127) at low pH.[58] Rapid photolysis and hydrolysis of omeprazole has also been reported in abiotic environments with a major deoxygenation transformation product (*m/z* 330.127).[59,60]

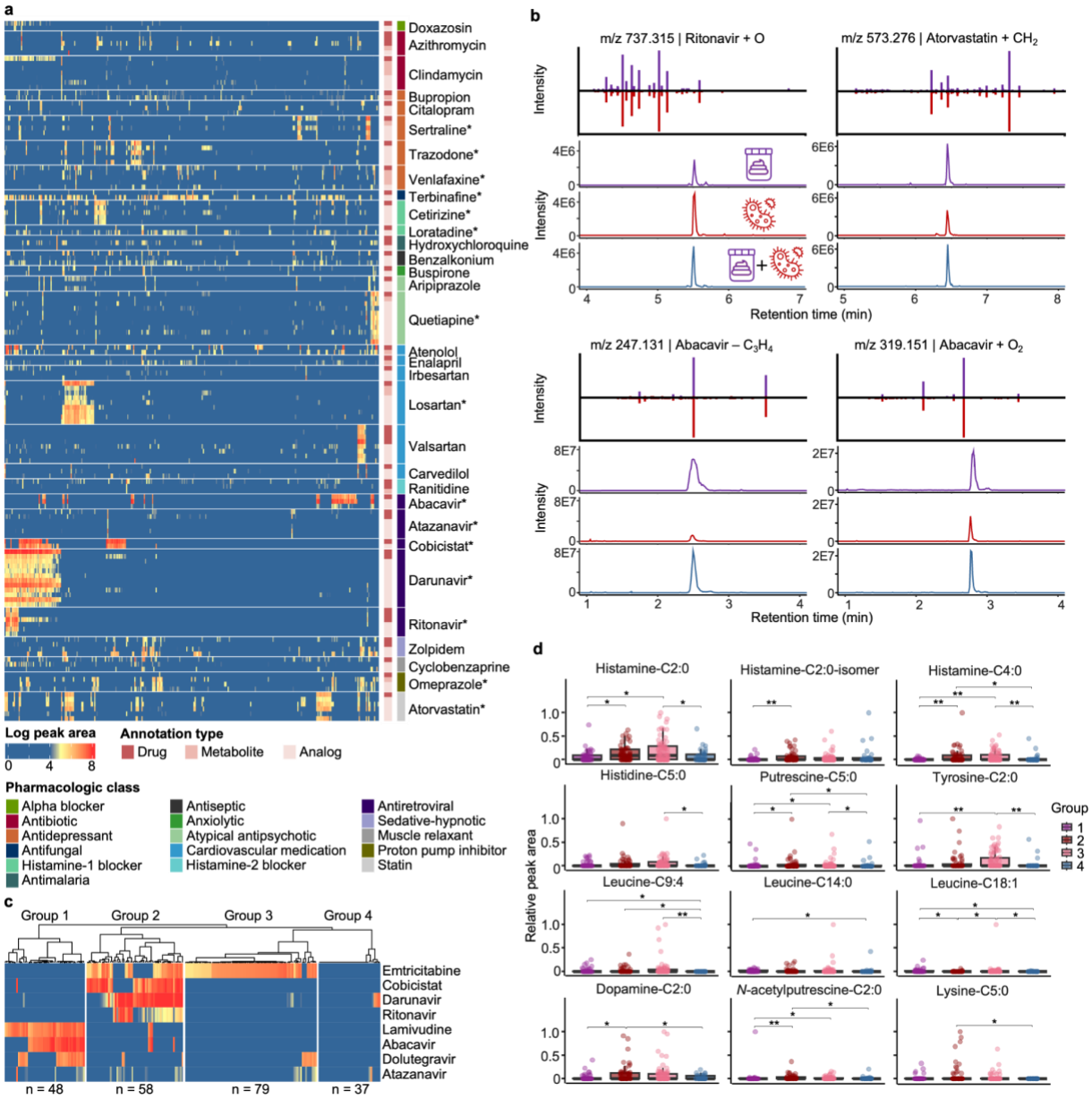

Figure 2. Drug exposures in the HIV Neurobehavioral Research Center (HNRC) cohort with connections to microbial metabolism and endogenous metabolites. From the HNRC cohort, 322 fecal samples were analyzed with 222 samples from people with HIV and 100 samples from people without HIV. **a,** Peak area visualization of drugs detected with metabolites and analogs. Each column represents one sample and each row represents one drug annotation. Drug annotations were grouped based on the parent drugs and separated by gap spaces. Drug annotations were denoted based on their types (as drug, drug metabolites, or drug analogs) and the pharmacologic classes of the parent drugs. All annotated ion/adduct forms of the parent drugs were visualized, leading to multiple rows of parent annotations for some drugs. Asterisks on the drug name mark parent drug annotations confirmed with commercial standards based on retention time and MS/MS spectral matches. Raw peak areas were log-transformed. **b,** Retention time and MS/MS spectra mirror matches for drug analogs observed in both the fecal samples and the drug microbial incubations. Purple traces represent the fecal samples, while red traces represent the drug microbial

incubation. Blue traces represent mixtures of the fecal samples and the microbial incubations at 1:1 volume ratio. The atomic changes of the drug analogs were based on $[M+H]^+$ ion of the parent drug. **c,** Hierarchical clustering of the samples from people with HIV (n = 222) based on detected antiretroviral drugs (ARV). Each row represents one detected ARV, with peak areas summed for the drug, metabolite, and analog detections followed by log-transformation (visualized with the same color scale as panel a). ARVs detected in <10% of samples are not shown. Each column represents one sample, clustered into four groups by hierarchical clustering with Ward's linkage and Euclidean distance. **d,** Sample-to-sample peak areas of *N*-acyl lipids in people with HIV, separated by the clusters derived from ARV detections shown in panel c. For each compound, the peak area in each sample was standardized to the maximum value observed across all samples. A non-parametric Kruskal-Wallis test followed by pairwise Wilcoxon test and Benjamini-Hochberg correction for multiple comparisons were performed. P values: * < 0.05; ** < 0.01. Boxplots showcase the median value, first (lower) and third (upper) quartiles, and whiskers indicate the error range as 1.5 times the interquartile range.

The GNPS Drug Library can enable stratification based on drug profiles, which facilitates discovery of connections between drug exposures and endogenous metabolites. *N*-acyl lipids are a class of signaling molecules made by host-associated microbiota[61] that play important roles in the immune system,[62] memory function,[63] and insulin regulation of the human body.[64–66] Our recent ongoing work found that the levels of histamine *N*-acyl lipids differed by HIV serostatus. Specifically, we observed higher levels of histamine-C2:0, histamine-C3:0, and histamine-C6:0 in people living with HIV than people without HIV.[67] To investigate whether these differences were related to drug exposures, we further stratified samples in this dataset by their ARV exposure profiles. As expected, ARV exposure profiles clearly separated based on the HIV serostatus (**Figure S6a**). High intensities of diverse ARVs were observed in fecal samples from people with HIV, while ARVs were only occasionally observed in people without HIV with low intensities. ARVs observed in people without HIV include tenofovir, maraviroc, atazanavir, and raltegravir, which are common ARVs used for prophylaxis (**Figure S6a**).[68,69] To control for the HIV serostatus and investigate the effects of ARV exposure, we excluded samples from people without HIV and stratified the people with HIV (n = 222) based on their ARV co-occurrences. Four distinct ARV exposure groups were observed based on hierarchical clustering that agreed well with the different combination antiretroviral therapy (cART) regimens commonly used (**Figure 2c**). For example, Group 1 (n = 48), characterized by lamivudine, abacavir, and dolutegravir exposures, corresponded to the dolutegravir/abacavir/lamivudine treatment regimen.[70] Group 2 (n = 58) with emtricitabine, darunavir, ritonavir, and cobicistat exposures, agreed with the darunavir/ritonavir regimen[71] and the darunavir/cobicistat/emtricitabine/tenofovir regimen.[72] Group 3 (n = 79), characterized by emtricitabine and dolutegravir exposures, may be related to the dolutegravir/emtricitabine/tenofovir treatment regimen (Group 2).[73] Group 4 (n = 37) were without apparent ARV exposures, possibly due to poor adherence, severe comorbidities, HIV elite control, or ARVs not included in the GNPS Drug Library or not amenable with LC-MS/MS detections (**Figure 2c**). Notably, we observed that the levels of histamine-C2:0 previously associated with HIV serostatus,[67] along with the levels of eleven other *N*-acyl lipids, were significantly different in the four ARV exposure groups (Kruskal-Wallis test, p-value < 0.05; **Figure 2d**). This suggests that exposure to different classes of ARV among people with HIV, in addition to HIV serostatus itself, might contribute to changed levels of these *N*-acyl lipids. We emphasize that these patterns could not have been revealed without the empirical drug readouts from untargeted metabolites.

Clinical research records may not document exposures to individual drugs and often do not provide quantitative information on the exposure levels. For example, metadata for the HNRC cohort on current ARV usage, which is based on self reports, documented drug usage as "ARV-naïve" (never received ARV), "no ARV" (no current ARV use), "non-HAART" (currently using less than three ARVs), and "HAART" (currently using three or more ARVs). Based on these classifications, no significant differences were observed for the 52 *N*-acyl lipids detected in these samples (**Figure S6b**). Without the empirical drug record readout, enabled by the GNPS Drug Library, the effects of drugs on microbial *N*-acyl lipid levels would be overlooked.

We anticipate the GNPS Drug Library to play a key role in precision medicine by enhancing our understanding of the effects of drugs across a wide range of phenotypes, including endogenous metabolism, gut and skin microbiome, pharmacokinetics, and drug-drug interactions. The empirical drug records from the GNPS Drug Library can enhance the clinical metadata by providing sample-to-sample comparisons of the relative abundance of individual drugs, which can be flexibly summarized at multiple ontology levels depending on user-defined questions. The mass spectrometry community will play a key role in the evolution of this resource through the continued deposition of reference libraries and expansion of the public metabolomics datasets for analog searches. By harnessing the power of public data and data science-ready metadata, we can unlock opportunities to deepen our understanding of the intricate relationships between xenobiotic exposure and human biological systems.

It is important to understand that the use of the GNPS Drug Library holds certain limitations. The current library only supports MS/MS-based annotations to level 2/3 according to the 2007 Metabolomics Standards Initiative.[74] This generally means that spectra of drug isomers may be annotated as the drug. Key drugs with important clinical implications should be checked for retention time matching and be quantified with analytical standards should the scientific question warrant this. The GNPS Drug Library can only capture drugs that are detectable in the specific biological matrix of choice (e.g., brain samples and urine will likely have different drug exposure readouts) and drugs that are ionizable with the chosen mass spectrometry setup. When constructing the drug analog library, we designed the filters to retain analog spectra that can be as confidently linked to drug exposure as possible, at the likely cost of excluding true positives. For example, metabolism pathways specific to substructures infrequently captured or missing in our customized delta mass list will be excluded. Metabolites shared between drugs that cannot be connected by the applied structural similarity scores (the Tanimoto score)[75] will be lost. As this is an evolving resource, we encourage the community to not only add to, but also report any inconsistencies in the library and the metadata they may notice.

Methods

Reference spectra collection and curation

Reference spectra of drugs were collected from the GNPS spectral library[25] and the MS^n library using a Python script developed for the MS^nLib workflow.[26] Detailed steps are summarized in the **Supplementary Text 1**. Briefly, the chemical structures provided in the open-source libraries are first cleaned by removing salt forms and standardized with the ChEMBL structure pipeline Python package.[76] Based on this cleaned structure, other structural information is calculated, including the canonical and, if available, isomeric SMILES, InChI, and InChIKey strings. Database identifiers, e.g., DrugBank ID or ChEMBL ID, are searched in PubChem and

Unichem based on the InChIKey string. Those identifiers, the complete InChIKey or the first part of the InChIKey to remove stereochemistry, are then used to search in drug databases, including the Broad Institute Drug Repurposing Hub (downloaded July 2022) ,[28] DrugBank (downloaded August 2022),[16,17] and DrugCentral (downloaded October 2022),[18] and the database of bioactive molecules from ChEMBL with clinical phases 1-4 (downloaded March 2023).[29] Reference spectra with chemical structures contained in any of the databases above are retained in the GNPS Drug Library.

To increase the coverage for drug metabolites and derivatives, we searched for reference spectra whose name contained the full name of a drug in the abovementioned databases, followed by manual inspection to remove mismatches. For example, by searching for the name "venlafaxine" (a serotonin and norepinephrine reuptake inhibitor indicated for the treatment of depression), we collected reference spectra for five metabolites of venlafaxine, including "*N*-desmethylvenlafaxine", "*O*-desmethylvenlafaxine", "*N*,*O*-didesmethylvenlafaxine", "*N*,*N*-didesmethylvenlafaxine", and "venlafaxine *N*-oxide" from the GNPS Library.

Drug analog search and result filtering

Drug analog spectra were retrieved from the GNPS/MassIVE public repository based on MS/MS spectra alignment using two computational methods: repository-scale molecular networking[41] and fastMASST with analog search.[42,43] A repository-scale molecular network was constructed in our previous research[41] and annotated with the 103,209 reference spectra of drugs and drug metabolites collected in this study. Unannotated spectra directly linked to annotated drugs or drug metabolites were collected as tentative drug analogs. FastMASST (fast Mass Spectrometry Search Tool)[42,43] checked the similarity of the queried spectrum against all spectra in the public repository without the pre-constructed molecular network. Batch mode FastMASST search for the 103,209 reference spectra of drugs and drug metabolites was achieved in July 2023 with the customized Python scripts (https://github.com/robinschmid/microbe_masst), leveraging the Fast Search tool API.[77] The Python scripts returned analog matches as metabolomics universal spectrum identifiers (USIs),[78] connecting to spectra in the GNPS/MassIVE repository. Analog spectrum matches were accepted with a modified cosine score of 0.8 or higher and a minimum of 6 matching ions, with a search window of 150 Da below and 200 Da above the precursor mass.

Filtering of the analog matches was achieved using customized Python scripts with the following steps. *Step One,* analog matches for drugs with endogenous or food sources, based on the manually curated metadata, were excluded to avoid matches to endogenous metabolites. *Step Two,* analogs with mass offsets (precursor mass differences between the analogs and the drugs) unrepresented by common metabolism pathways were excluded. Specifically, mass offsets sourced from UNIMOD,[44] from a community-curated list of delta masses (**Table S2**), and from the Host Gut Microbiota Metabolism Xenobiotics Databases[45] were manually curated to only keep those relevant to drug metabolism (e.g., 14.02 Da, methylation; 176.03, glucuronidation) or mass spectrometry adducts (e.g., 17.03, ammonium adduct; see **Table S3** for the 156 mass offsets included). Analogs were excluded if the mass offsets were not in the customized list, or if the mass offsets occurred fewer than ten times. *Step Three,* analogs with matches to the GNPS Library were excluded to restrict the analog library to true unknowns. This is achieved through fastMASST searches of the GNPS Library (analog search off, minimum cosine score 0.7, minimum 5 matching ions); all USIs with library matches were excluded from the drug analogs.

Step Four, analogs matched to multiple drugs with dissimilar structures were excluded. Specifically, spectra of the remaining USIs from Step Three were downloaded using the metabolomics-usi API[78] and clustered by the falcon spectrum clustering tool (with default setting except an eps score of 0.05).[79] Singleton spectra were excluded. For clustered spectra connected to multiple drugs, Tanimoto similarities of the drugs need to be higher than 0.8 to accept the drug analogs.[75] *Step Five,* the drug analog libraries were tested against 10 datasets to remove edge cases where the analogs were detected with high frequency and cannot be confidently linked to drug exposure. The testing datasets include four datasets for human fecal samples (GNPS/MassIVE ID: MSV000080673, MSV000095418, MSV000094515, MSV000092833), two for human breast milk samples (MSV000091520, MSV000090877), one for human plasma (MSV000094395), one for human brain tissues (MSV000086415), and two for mouse tissues (MSV000091007, MSV000091363). Search against the GNPS Drug Library was performed using the Library Search workflow on GNPS with minimum cosine score of 0.7 and minimum 5 matching peaks.

To extend structural hypotheses for the drug analogs, we employed the newly developed ModiFinder,[32] which leverages the shifted MS/MS fragment peaks in the MS/MS alignment to predict the most likely location for the structural modifications. ModiFinder was run with mass tolerance of 40 ppm and fragmentation depth of 2. The GNPS job is available at: https://gnps2.org/result?task=f4bf3bab255c4bd8a4175433286d5f2c&viewname=modifinder_result&resultdisplay_type=task.

Metadata curation

The GNPS Drug Library provides ontology-based metadata on the chemical sources, therapeutic areas, pharmacologic classes, therapeutic indications, and mechanisms of action of the drugs. The chemical source information was first retrieved from the ChemFOnt database[47] through name matching. Synonyms of the drugs in the GNPS Drug Library were extracted from PubChem with the MS^n library Python scripts and searched against chemical names in the ChemFOnt database. The results were manually curated by three experts through parsing web pages and scientific literature to increase compound coverage (all drugs in the GNPS Drug Library have chemical source information) and improve accuracy and consistency. The therapeutic areas, mechanisms of action, and therapeutic indications of the drugs were extracted from the Broad Institute Drug Repurposing Hub[28] using the MS^n library Python scripts through InChIKey matches. The pharmacologic classes were extracted from the U.S. Food and Drug Administration (FDA) documents by synonym matching. The above database search returns the therapeutic areas, mechanisms of action, and therapeutic indications of 3,894 drugs (86,364 spectra) and the pharmacologic classes of 900 drugs (22,935 spectra). The extracted metadata were manually curated by expert pharmacologists to improve accuracy and fill in missing information to the extent possible, which increased the metadata coverage to 4,560 drugs (90,325 spectra). Sources used to curate the metadata included the Anatomical Therapeutic Chemical (ATC) Classification, National Institutes of Health (NIH), FDA, European Medicines of Agency (EMA), and DrugBank. In Particular, the WHO ATC codes were used in the manual curation of the pharmacologic class. The refined metadata can be illustrated with doxycycline, a tetracycline-class drug. Initially, it was classified under dental as a therapeutic area with periodontitis as the only indication. Following manual curation, doxycycline was reclassified under infectious diseases (therapeutic area) and Gram-positive and Gram-negative infections (therapeutic indication), recognizing its usage to

treat a variety of infections. For drug metabolites and analogs, metadata corresponding to the parent drug was used.

Reanalysis of public data in ReDU

The MGF spectra files of the GNPS Drug Library were imported into the fastMASST pipeline[42,43] and searched against the GNPS/MassIVE repository using the following parameters: 0.05 for both the precursor *m/z* tolerance and fragment *m/z* tolerance, 0.7 cosine threshold, 5 minimum matched peaks, and analog search off. In order to have an overview of the drug exposures among different disease groups, the fastMASST results were merged with ReDU ("Reanalysis of Data User Interface") controlled vocabulary metadata, which is an interface linking metadata with curated controlled vocabularies to public untargeted metabolomics data files.[48] This merged table was then filtered to contain only matches to human datasets ("9606|Homo sapiens" in the NCBITaxonomy column). After filtering, the unique combinations of disease status and body parts were extracted based on the "DOIDCommonName" and the "UBERONBodyPartName" column. Detected drugs were connected to their therapeutic areas using the GNPS Drug Library metadata, separating the "infectious disease" drugs into antibiotics, antifungals, and antivirals. The heatmap was generated to visualize exposure profiles based on disease status by calculating the numbers of samples with drugs in a certain therapeutic area normalized to the numbers of samples in each sample type with at least one drug detection. The heatmap was generated using the 'ComplexHeatmap' package in R (v 2.20.0) with script provided in the code availability section.

Reanalysis of the American Gut Project data

In order to overview drug exposure profiles among ages and genders, the mzML files from the public metabolomics dataset of the American Gut Project (MSV000080673) were re-processed with feature extraction in MZmine 3.[80] The data were originally acquired with LC-qTOF-MS/MS and the following MZmine settings were used: For mass detection, the noise levels were 1E3 for MS1 detection and 5E1 for MS2 detection. For chromatogram building, the mass tolerance was set as 0.0050 *m/z* or 30 ppm, the minimum consecutive scans as 5, and the minimum height as 3E3. For local minimum search for chromatographic deconvolution, minimum search range was 0.1 min, minimum ratio of peak top to side was 2, and maximum peak duration was 2.0 min. The peaks were de-isotoped within 3 ppm *m/z* and 0.08 minutes retention time tolerances, then aligned with 0.0050 *m/z* or 30 ppm mass tolerance and 0.3 minutes retention time tolerance. The feature list was filtered for minimum detections in 2 samples and exported as a feature quantification table (.csv) and an MGF spectra file. The feature quantification table was subsequently filtered with R scripts to remove features with average peak areas in samples lower than 3-fold of those in blanks. For the remaining features, peak areas in samples were considered detected only if they were higher than 3-fold of the average peak areas in blanks.

Annotations were performed with the GNPS Drug Library using the Feature-Based Molecular Networking (FBMN) workflow on GNPS.[81] The spectra were filtered by removing MS/MS fragments within ±17 Da of the precursor *m/z* and to only keep the six most intense fragments in ±50 Da window. The spectra were searched against the GNPS Drug Library with precursor and fragment ion mass tolerances of 0.01 Da, a cosine score threshold of 0.7 and minimum 2 matched peaks. The GNPS job is available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=649a6b4bfc5a4b1d9131b3ff23c1ae12. Annotations with cosine score lower than 0.9 and matched peaks fewer than 5 were removed in

downstream analysis. The remaining annotations were manually inspected and were accepted only if two or more major fragment ions were matched.

Demographic information of the samples were collected from ReDU controlled vocabulary metadata, including "AgeInYears" column for age, "BiologicalSex" column for sex, "Country" for country of residence, and "LatitudeandLongitude" for geographic coordinates. Samples from the intensive care unit (ICU) microbiome pilot (Qiita study 2136)[82] were excluded from the analysis because their drug exposure does not resemble the general population (n = 93). In cases where individuals had more than 1 sample (n = 47), only the first collected sample was included in the analysis. Only individuals from the US, Europe, and Australia were included in the analysis, due to few samples from other regions (n = 50). Annotations were filtered based on the exposure source to exclude drugs with endogenous or diet-derived sources. Drug annotations were further grouped based on unique parent compound names, and the number of drugs detected per individual were counted to create the world map visualization (**Figure S3b**). Geographic coordinates were available for n = 1,903 individuals. The world map was generated using the "rworldmap" package in R with scripts provided in the code availability section. For the visualization based on pharmacologic classes (**Figure 1g-h**), drug annotations were further grouped based on the pharmacologic class metadata provided in the GNPS Drug Library. Each pharmacologic class was counted only once per individual when multiple drugs under the same class were detected. Age groups were defined in 10-years intervals (nine intervals for age 0-90 years old), and individuals with no information on age and sex (n = 148) and individuals >90 years (n = 2) were excluded from the analysis. To adjust for the varying number of samples across age and sex, normalization was performed by dividing the number of observations by the total number of samples within each age group and sex.

Extraction and LC-MS/MS profiling of human feces with HIV infection

Sample extraction. The UCSD Human Research Protections Program (irb.ucsd.edu) approved all study procedures before the initiation of the study and all participants provided written informed consent. The stool samples were prepared with a recently developed automatic pipeline for simultaneous metagenomics-metabolomics extractions.[83] Specifically for the metabolite extraction, swabs were added into Matrix Tubes (ThermoFisher Scientific, MA, USA) containing 400 µL of 95% (v/v) ethanol and capped with the automated instrument Capit-All (ThermoFisher Scientific, MA, USA). The tubes were shaken for 2 minutes (1,200 rpm, the SpexMiniG plate shaker) followed by centrifugation for 5 minutes (2,700 g). The supernatant was then transferred to a deep well plate using an 8-channel pipette, concentrated to dryness with a vacuum centrifuge concentrator (room temperature; ~5 hours) and stored at -80 °C. Prior to instrumental analysis, the plates were redissolved (10 minute sonication) in 200 µL of 50% acetonitrile (v/v) with 100 µg/L sulfadimethoxine as the internal standard. Plates were centrifuged at 450 g for 10 minutes, and 150 µL supernatants were collected for instrumental analysis.

Instrumental analysis. The fecal extracts were injected (5 µL) into a Vanquish ultra-high-performance liquid chromatography (UHPLC) system coupled to a Q Exactive quadrupole orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA). A Kinetex polar C18 column (150 × 2.1 mm2, 2.6 µm particle size, 100 A pore size; Phenomenex, Torrance) was employed with a SecurityGuard C18 column (2.1 mm ID) at 30 °C column temperature. The mobile phases (0.5 mL/min) were 0.1% formic acid in both water (A) and ACN (B) with the following gradient: 0-1 min 5 % B, 1-7 min 5-99 % B, 7-8 min 99 % B, 8-8.5 min 99-5% B, 8-10 min 5% B. The mass

spectrometer was operated in positive heated electrospray ionization with the following parameters: sheath gas flow, 53 AU; auxiliary gas flow, 14 AU; sweep gas flow, 3 AU; auxiliary gas temperature, 400 °C; spray voltage, 3.5 kV; inlet capillary temperature, 269 °C; S-lens level, 50 V. MS1 scan was performed at *m/z* 100-1500 with the following parameters: resolution, 35,000 at *m/z* 200; maximum ion injection time, 100 ms; automatic gain control (AGC) target, 5.0E4. Up to 5 MS/MS spectra per MS1 scan were recorded under the data-dependent mode with the following parameters: resolution, 17,500 at *m/z* 200; maximum ion injection time, 100 ms; AGC target, 5.0E4; MS/MS precursor isolation window, *m/z* 3; isolation offset, *m/z* 0.5; normalized collision energy, a stepwise increase from 20 to 30 to 40 %; minimum AGC for MS/MS spectrum, 5.0E3; apex trigger, 2 to 15 s; dynamic precursor exclusion, 10 s.

Feature extraction and data analysis. The raw spectra were converted to mzML files using MSconvert (ProteoWizard)[84] followed by feature extraction using MZmine 4[80] with the following parameters: For mass detection, the noise factor was 3.5 for MS1 and 2.5 for MS2. For chromatogram building, the mass tolerance was set as 0.01 *m/z* or 10 ppm, the minimum consecutive scans as 3, and the minimum height as 5E2. Chromatograms were smoothed with the Savitzky Golay algorithm, followed by local minimum search for chromatographic deconvolution with minimum search range of 0.2 min, minimum ratio of peak top to side of 2, and maximum peak duration of 0.5 min. The peaks were de-isotoped within 5 ppm *m/z* and 0.05 minutes retention time tolerances, aligned with 0.01 *m/z* or 20 ppm mass tolerance and 0.3 minutes retention time tolerance, then gap-filled with 20 ppm *m/z* tolerance and 0.3 minutes retention time tolerance. The feature list was blank-subtracted with 300% fold change increases compared with maximum blank values and 30 minimum number of detection in blanks. The feature list was then exported as a feature quantification table (.csv) and an MGF spectra file and used without post-filtering.

The features were annotated for drugs and drug analogs using the GNPS Drug Library with the same procedure as for the American Gut Project data. The GNPS job is available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=90b6edcfb8fc43dab54d2d7951a3291a. Peak areas of annotated drugs lower than 1E4 were replaced by zero to avoid false detections caused by instrument noises. Representative drug annotations, including all annotated ARVs and drugs used in the microbial incubations, were validated by analytical standards (**Figure S7**).The features were annotated for *N*-acyl lipids using a recently developed a spectral library for 851 *N*-acyl lipids[67] with the following parameters for the FBMN workflow: 0.02 Da precursor and fragment ion mass tolerance and library search with a cosine score threshold of 0.7 and minimum 4 matched peaks. The GNPS job is available at https://gnps2.org/status?task=ee34ee95908749dd81ee9a62fbdac98e. The heatmaps to demonstrate drug annotations were generated using the 'ComplexHeatmap' package in R (v 2.20.0). People with HIV were clustered into four groups based on their ARVs profiles by hierarchical clustering under the Ward's linkage method and Euclidean distance matrix. Boxplots representing the peak areas of *N*-acyl lipids in different drug exposure groups were generated using the 'ggplot2' package in R (v 3.5.1), with R scripts provided in the code availability section.

Synthetic microbial community culture for drug metabolism

Chemicals. Drugs for bacterial metabolism screening were purchased from Sigma-Aldrich (St. Louis, MO, USA; see detailed list in **Table S4**). All solvents used were Optima LC-MS grade from Fisher Scientific (Pittsburgh, PA, USA). We note that quinine and quinidine were included in

the drug metabolism assays because they were initially annotated based on MS/MS spectra matches (level 2/3 according to the 2007 Metabolomics Standards Initiative)[74]. After we acquired the standards, their retention times do not match the peaks in fecal samples. Therefore, quinine and quinidine were not considered in downstream analysis.

Synthetic microbial community culture. The 111 bacterial strains used in this study are a selection of the human gut microbiota (hCom) previously published[57] that was optimized in the Zengler lab to accomplish uniform growth (**Table S5**). A normalized working stock solution at OD_{600} = 0.01 was created for the 111 bacterial strains and stored at -80 ℃ until needed. For culturing, six groups were created and include the full hCom, the hCom-reduced (*Coprococcus comes* and *Coprococcus eutactus* were omitted due to their dominance in the community), fast growers, medium-fast growers, medium-slow growers, and slow growers (**Table S5**). To create the hCom and hCom-reduced communities, the normalized stock of a bacteria was thawed, and a specific volume was taken and combined to generate the groups. To create the hCom-reduced, each bacteria was added based on their groups (fast-growers, 0.5 µL; medium-fast growers, 5 µL; medium-slow, 50 µL (not including the two species of *Coprococcus*); and slow growers, 500 µL). *Coprococcus comes* and *Coprococcus eutactus* were added at 0.5 µL to finalize the assembly of the full hCom. To create the other group communities, 500 µL of each bacteria of each group were combined (**Table S5**). The groups were then diluted 1/100 at a final OD_{600} of 0.0001 and cultured anaerobically in BHI medium (**Table S6**) at 37 ℃ in an anaerobic chamber (10% CO_2, 7.5% H_2, 82.5% N_2).

Drug metabolism assays. A 10 mM stock solution was created for each drug and diluted to 200 µM with H_2O. Drugs were combined to create four different mixtures for the screening assay (mix1: sertraline, quinidine, venlafaxine; mix2: cetirizine, losartan, quetiapine, ritonavir; mix3: darunavir, quinine, atorvastatin, omeprazole; mix4: loratadine, terbinafine, trazodone, abacavir). Each drug was added to a final concentration of 2 µM (2 µL spike) after 24 h of bacterial acclimation. Cultures were immediately extracted after drug addition as the 0 h samples. Cultures from the fast and medium-fast groups were extracted after 48 h of growth and the cultures from the remaining groups were stopped at 72 h. Medium controls were included by adding drug mixtures to the BHI medium without microbial communities and were included at both 0 h and 72 h. All conditions were performed in triplicates.

Extraction of drugs and metabolites for LC-MS analysis. Bacterial cultures were extracted using 600 µL of pre-chilled 50% (v/v) MeOH, followed by an overnight incubation at 4℃. Samples were dried using a CentriVap and stored at -80℃ until LC-MS/MS analysis. Samples were resuspended in 200 µL of pre-chilled 50% (v/v) MeOH with 100 µg/L sulfadimethoxine as an internal standard, incubated at -20℃ overnight, centrifuged at 15,000 x g for 10 min, and 150 µL of the supernatant was transferred to 96 shallow well-plate for LC-MS analysis. Samples were analyzed with the same LC-MS/MS method for the HIV feces samples, except the following parameters: additional LC washing gradient after sample analysis, 8.5-9 min 5% B, 9-9.5 min 5-99% B, 9.5-10.5 min 99% B, 10.5-11 min 99-5% B, 11-12.5 min 5% B; MS1 scan range, *m/z* 100-100; AGC target, 1E6; dd-MS2 AGC target, 5.0E5; MS/MS precursor isolation window, *m/z* 1.

The raw spectra were converted to mzML files followed by feature extraction using MZmine 4 with the following parameters: For mass detection, the noise factor was 3.0 for MS1 and 2.0 for MS2. For chromatogram building, the mass tolerance was set as 0.002 *m/z* or 10 ppm, the minimum consecutive scans as 4, and the minimum height as 1.5E5. Chromatograms were

smoothed with the Savitzky Golay algorithm, followed by local minimum search for chromatographic deconvolution with minimum search range of 0.05 min, minimum ratio of peak top to side of 2, and maximum peak duration of 1.5 min. The peaks were de-isotoped within 3 ppm *m/z* and 0.04 minutes retention time tolerances, aligned with 0.002 *m/z* or 10 ppm mass tolerance and 0.07 minutes retention time tolerance, then gap-filled with 10 ppm *m/z* tolerance and 0.05 minutes retention time tolerance.

The features were annotated for drugs and drug analogs using the GNPS Drug Library with the same procedure as for the American Gut Project data. The GNPS job is available at: https://gnps2.org/status?task=128c85f310844f2d90d71bcf1c107afb.

Co-migration analysis with fecal extracts. Four samples from the bacterial cultures were analyzed again with selected fecal samples from the HIV infection cohort to confirm the retention time and MS/MS spectral matching of the drug analogs. One medium-slow grower at 72 hours was used to confirm the ritonavir analog at *m/z* 737.31, one medium-fast grower at 72 hours for the atorvastatin analog at *m/z* 573.28, one medium-fast grower at 72 hours for abacavir analogs at *m/z* 247.13 and 319.15, and one hCom-reduced culture at 0 hour for the omeprazole analog at *m/z* 330.13. Fecal samples with the highest analog intensities were used. Extracts of the bacterial cultures or feces were first diluted to similar concentrations with 50% (v/v) methanol based on peak areas of the drug analogs in the initial sample analysis, and then were mixed at 1:1 volume ratio. The selected bacterial culture samples, fecal samples, and mixture samples were co-analyzed using the same LC-MS/MS method for the bacterial cultures.

Statistical Analysis.

A chi-square test was performed to compare the number of drugs detected in fecal samples from different countries (**Figure S3b**). Kruskal-Wallis tests followed by a pairwise Wilcoxon test were performed with Benjamini-Hochberg correction to compare the levels of *N*-acyl lipids in the four drug exposure groups (**Figure 2d**, **Figure S6d**). Adjusted p-values <0.05 were considered as statistically significant. Pairwise Wilcoxon tests were performed to compare the drug analog intensities in bacterial cultures at 0 hour and 72 hours (**Figure S5**). Due to the small sample size (n = 3 in each group), p-values <0.1 were considered as statistically significant.

Acknowledgements

This project was enabled in part by the Alzheimer's Gut Microbiome Project (AGMP) and the Data Infrastructure and Molecular Atlas for AD: Connection Exposome, Gut Microbiome, and Metabolome supplement funded wholly or in part by the following grants thereto: 1U19AG063744 and 3U19AG063744-04S1 and awarded to Dr. Kaddurah-Daouk at Duke University in partnership with multiple academic institutions. As such, the investigators within the AGMP and the Exposome Supplement, not listed specifically in this publication's author's list, provided data along with its pre-processing and prepared it for analysis, but did not participate in analysis or writing of this manuscript. A listing of AGMP Investigators can be found at https://alzheimergut.org/meet-the-team/. A complete listing of ADMC investigators can be found at: https://sites.duke.edu/adnimetab/team/. We also thank the support by NIH for the Maternal and Pediatric Precision in Therapeutics project P50HD106463, the development of tools for structure elucidation R01DK136117, the Collaborative Microbial Metabolite Center U24DK133658. The HIV Neurobehavioral Research Center (HNRC) is supported by Center award P30MH062512 from NIMH. This research was supported in part by the Intramural

Research Program of the NIH, National Institute of Environmental Health Sciences (ZIC ES103363). C.B. was supported by the Czech Academy of Sciences PPLZ fellowship number L200552251. V.C.L is supported by Fonds de recherche du Québec - Santé (FRQS) Postdoctoral fellowship (335368). N.E.A was supported in part by the National Center for Complementary and Integrative Health of the NIH under award number F32AT011475. A.M.C.-R. and P.C.D. were supported by the Gordon and Betty Moore Foundation grant GBMF12120. M.R. was supported by the NIH grant R37 AI126277. T.P. was supported by the Czech Science Foundation (GA CR) grant 21-11563M and by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 891397.

Disclosures

R.S.: R.S. is a co-founder of mzio GmbH.

D.M.: D.M. is a consultant for BiomeSense, Inc., has equity and receives income. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

R.K.-D.: R.K.-D. is an inventor on a series of patents on use of metabolomics for the diagnosis and treatment of CNS diseases and holds equity in Metabolon Inc., Chymia LLC and PsyProtix.

M.W.: M.W. is a co-founder of Ometa Labs LLC

T.P.: T.P. is a co-founder of mzio GmbH.

R.K.: R.K. is a scientific advisory board member, and consultant for BiomeSense, Inc., has equity and receives income. He is a scientific advisory board member and has equity in GenCirq. He is a consultant for DayTwo, and receives income. He has equity in and acts as a consultant for Cybele. He is a co-founder of Biota, Inc., and has equity. He is a cofounder of Micronoma, and has equity and is a scientific advisory board member. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

S.M.T.: S.M.T. receives research funding from Veloxis Pharmaceuticals.

P.C.D.: P.C.D. is a scientific advisor and holds equity in 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

H.N.Z., C.B., P.C.D. conceptualized the method. H.N.Z, C.B., R.S., T.P. developed the MS/MS library for drugs. H.N.Z., W.B., R.T., R.S. developed the MS/MS library for drug analogs. H.N.Z., S.M., H.S., P.R., curated exposure source metadata. K.E.K. curated pharmacological metadata. H.N.Z., R.T., Y.E.A., H.M.-R., N.E.A., A.M.C.-R., P.W.P.G., S.M., I.M., A.C., S.P.T., S.Z., curated drug name match results. H.N.Z., K.E.K., S.L., H.M.-R., L.K., S.X., performed data analyses. Y.E.A., M.S.A., C.W., A.K.J., D.M., helped with data interpretation. V.C.-L., H.N.Z., L.C., C.W., M.R., K.Z. performed microbial incubation experiments. D.H.R., L.X. contributed MS/MS reference spectra. M.S.A., R.J.E., D.F., R.H., J.I., S.L., D.J.M. developed the clinical cohort of human immunodeficiency virus (HIV) infection. M.R.Z.S., M.W. performed ModiFinder analysis. S.G., D.S.W. provided support on exposure source annotation. R.K.-D. supervised the consortium providing access to the Alzheimer's disease cohort and acquired funding. R.K. supervised sample handling and DNA data acquisition for the Alzheimer's disease, HIV and American Gut Project

cohorts, and acquired funding. H.N.Z., K.E.K., C.B., P.C.D. drafted the manuscript. P.C.D., S.M.T. acquired funding and supervised this project. All authors reviewed and edited the manuscript.

Data availability

The MGF spectral files for the GNPS Drug Library and the associated metadata of controlled vocabularies (.csv) can be downloaded from Zenodo archive under doi: 10.5281/zenodo.13892289. The downloaded MGF spectral files can be added to personal GNPS folders and used directly for library matching. Data used to profile drug exposures at age and gender groups are publicly available in GNPS/MassIVE under the accession number MSV000080673. Data for fecal samples from the HNRC cohort are available at MSV000092833. Data for the drug bacterial cultures are available at MSV000095331. Data for HNRC fecal samples analyzed with the bacterial cultures are available at MSV000096012. Data for co-migration of the bacterial cultures and fecal samples are available at MSV000096013.

Code availability

The code used to query reference spectra of drugs is available on GitHub under the MS^n library project[26] (https://github.com/corinnabrungs/msn_tree_library). The code used to filter the drug analog matches is provided on GitHub (https://github.com/ninahaoqizhao/Manuscript_GNPS_Drug_Library). The code used for dataset analysis can be found on GitHub (https://github.com/ninahaoqizhao/Manuscript_GNPS_Drug_Library and https://github.com/kinekvitne/manuscript_drug_library).

Supplementary Figure

Supplementary Figure S1. Additional examples of structural modification sites of the drug analogs predicted by ModiFinder; Supplementary Figure S2. Overview of the ontology-based drug metadata based on the numbers of reference spectra. Supplementary Figure S3. Drug exposure profiles among different disease types and geolocations by re-analyzing public metabolomics data. Supplementary Figure S4. Drug analog annotations in fecal samples from people with human immunodeficiency virus. Supplementary Figure S5. Drug analogs observed in human fecal samples can be produced by microbial metabolism. Supplementary Figure S6. Comparison of sample clustering based on empirical drug records from the GNPS Drug Library or on clinical metadata. Supplementary Figure S7. Retention time and MS/MS spectra mirror matches for drugs observed in the HNRC cohort with analytical standards.

Supplementary Table

Supplementary Table S1. Community-curated list of delta mass interpretation; Supplementary Table S2. Delta masses accepted in the drug analog library; Supplementary Table S3. Percentage of drug analogs demonstrating co-occurrence with the parent drugs in MASST search, separated by delta masses; Supplementary Table S4. Source of drugs used in synthetic microbial community incubation; Supplementary Table S5. Bacterial strains used in the six synthetic microbial communities; Supplementary Table S6. Composition of the BHI medium for anaerobic microbial cultures.

References

1. Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).
2. National Center for Health Statistics (U.S.), Health, United States, 2019 (U.S. Centers for Disease Control and Prevention, 2021).
3. Rappaport, S. M., Barupal, D. K., Wishart, D., Vineis, P. & Scalbert, A. The blood exposome and its role in discovering causes of disease. *Environ. Health Perspect.* **122**, 769–774 (2014).
4. de la Cuesta-Zuluaga, J., Boldt, L. & Maier, L. Response, resistance, and recovery of gut bacteria to human-targeted drug exposure. *Cell Host Microbe* **32**, 786–793 (2024).
5. Verdegaal, A. A. & Goodman, A. L. Integrating the gut microbiome and pharmacology. *Sci. Transl. Med.* **16**, eadg8357 (2024).
6. Vich Vila, A. et al. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nat. Commun.* **11**, 362 (2020).
7. Maier, L. et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
8. Sinnott, S.-J., Polinski, J. M., Byrne, S. & Gagne, J. J. Measuring drug exposure: concordance between defined daily dose and days' supply depended on drug class. *J. Clin. Epidemiol.* **69**, 107–113 (2016).
9. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **9**, 211–217 (2016).
10. Gong, Y. et al. Over-the-counter antibiotic sales in community and online pharmacies, China. *Bull. World Health Organ.* **98**, 449–457 (2020).
11. Mackey, T. K. et al. Multifactor quality and safety analysis of antimicrobial drugs sold by online pharmacies that do not require a prescription: Multiphase observational, content analysis, and product evaluation study. *JMIR Public Health Surveill.* **8**, e41834 (2022).
12. Essigmann, H. T. et al. Epidemiology of antibiotic use and drivers of cross-border procurement in a Mexican American border community. *Front. Public Health* **10**, (2022).
13. Walmsley, B. et al. The PrEP You Want: A web-based survey of online cross-border shopping for HIV prophylaxis medications. *J. Med. Internet Res.* **21**, e12076 (2019).
14. Vauquelin, G. & Charlton, S. J. Long-lasting target binding and rebinding as mechanisms to prolong in vivo drug action. *Br. J. Pharmacol.* **161**, 488–508 (2010).
15. Copeland, R. A. The drug–target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.* **15**, 87–95 (2016).
16. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
17. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
18. Ursu, O. et al. DrugCentral: online drug compendium. *Nucleic Acids Res.* **45**, D932–D939 (2017).
19. DailyMed. https://dailymed.nlm.nih.gov/dailymed/index.cfm.
20. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
21. Wan, M. et al. TnT-LLM: Text mining at scale with large language models. Preprint at https://doi.org/10.48550/arXiv.2403.12173 (2024).

22. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
23. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 1–8 (2023).
24. Prakash, C., Shaffer, C. L. & Nedderman, A. Analytical strategies for identifying drug metabolites. *Mass Spectrom. Rev.* **26**, 340–369 (2007).
25. 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).
26. Brungs, C. et al. Efficient generation of open multi-stage fragmentation mass spectral libraries. Preprint at https://doi.org/10.26434/chemrxiv-2024-l1tqh (2024).
27. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
28. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
29. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
30. Nahler, G. Anatomical therapeutic chemical classification system (ATC). in *Dictionary of Pharmaceutical Medicine* 8–8 (Springer, Vienna, 2009).
31. Ross, D. H., Seguin, R. P., Krinsky, A. M. & Xu, L. High-throughput measurement and machine learning-based prediction of collision cross sections for drugs and drug metabolites. *J. Am. Soc. Mass Spectrom.* **33**, 1061–1072 (2022).
32. Shahneh, M. R. Z. et al. ModiFinder: Tandem mass spectral alignment enables structural modification site localization. *J. Am. Soc. Mass Spectrom.* (2024).
33. McDonald, D. et al. American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, e00031-18 (2018).
34. Products - Data Briefs - Number 347 - August 2019. https://www.cdc.gov/nchs/products/databriefs/db347.htm (2019).
35. Helms, P. J., Ekins Daukes, S., Taylor, M. W., Simpson, C. R. & McLay, J. S. Utility of routinely acquired primary care data for paediatric disease epidemiology and pharmacoepidemiology. *Br. J. Clin. Pharmacol.* **59**, 684–690 (2005).
36. Isacson, D. & Bingefors, K. Epidemiology of analgesic use: a gender perspective. *Eur. J. Anaesthesiol. Suppl.* **26**, 5–15 (2002).
37. Anthony, M. et al. Gender and age differences in medications dispensed from a national chain drugstore. *J. Womens Health* **17**, 735–743 (2008).
38. Quinn, R. A. et al. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends Pharmacol. Sci.* **38**, 143–154 (2017).
39. Chen, L. et al. Metabolite discovery through global annotation of untargeted metabolomics data. *Nat. Methods* **18**, 1377–1385 (2021).
40. Yu, N. et al. Nontarget discovery of antimicrobial transformation products in wastewater based on molecular networks. *Environ. Sci. Technol.* **57**, 8335–8346 (2023).
41. Bittremieux, W. et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nat. Commun.* **14**, 8488 (2023).
42. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).

43. Mongia, M. et al. Fast mass spectrometry search and clustering of untargeted metabolomics data. *Nat. Biotechnol.* 1–6 (2024).
44. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536 (2004).
45. Kolodnitsky, A. S., Ionov, N. S., Rudik, A. V., Filimonov, D. A. & Poroikov, Vladimir. V. HGMMX: Host Gut Microbiota Metabolism Xenobiotics Database. *J. Chem. Inf. Model.* **63**, 6463–6468 (2023).
46. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **627**, 71–81 (2008).
47. Wishart, D. S. et al. ChemFOnt: the chemical functional ontology resource. *Nucleic Acids Res.* **51**, D1220–D1229 (2023).
48. Jarmusch, A. K. et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
49. Audi, S. et al. The 'top 100' drugs and classes in England: an updated 'starter formulary' for trainee prescribers. *Br. J. Clin. Pharmacol.* **84**, 2562–2571 (2018).
50. Fuentes, A. V., Pineda, M. D. & Venkata, K. C. N. Comprehension of top 200 prescribed drugs in the US as a resource for pharmacy teaching, training and practice. *Pharm. J. Pharm. Educ. Pract.* **6**, 43 (2018).
51. Wylie, C. E., Daniels, B., Brett, J., Pearson, S.-A. & Buckley, N. A. A national study on prescribed medicine use in Australia on a typical day. *Pharmacoepidemiol. Drug Saf.* **29**, 1046–1053 (2020).
52. Winston, A. & Spudich, S. Cognitive disorders in people living with HIV. *Lancet HIV* **7**, e504–e513 (2020).
53. Arseniou, S., Arvaniti, A. & Samakouri, M. HIV infection and depression. *Psychiatry Clin. Neurosci.* **68**, 96–109 (2014).
54. Donati, K. de G., Rabagliati, R., Iacoviello, L. & Cauda, R. HIV infection, HAART, and endothelial adhesion molecules: current perspectives. *Lancet Infect. Dis.* **4**, 213–222 (2004).
55. Heaton, R. K. et al. Twelve-year neurocognitive decline in HIV is associated with comorbidities, not age: a CHARTER study. *Brain J. Neurol.* **146**, 1121–1131 (2023).
56. Schmid, R. et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat. Commun.* **12**, 3832 (2021).
57. Cheng, A. G. et al. Design, construction, and in vivo augmentation of a complex gut microbiome. *Cell* **185**, 3617-3636.e19 (2022).
58. Olbe, L., Carlsson, E. & Lindberg, P. A proton-pump inhibitor expedition: the case histories of omeprazole and esomeprazole. *Nat. Rev. Drug Discov.* **2**, 132–139 (2003).
59. Boix, C., Ibáñez, M., Sancho, J. V., Niessen, W. M. A. & Hernández, F. Investigating the presence of omeprazole in waters by liquid chromatography coupled to low and high resolution mass spectrometry: degradation experiments. *J. Mass Spectrom.* **48**, 1091–1100 (2013).
60. Mathew, M., Gupta, V. D. & Bailey, R. E. Stability of omeprazole solutions at various ph values as determined by high-performance liquid chromatography. *Drug Dev. Ind. Pharm.* **21**, 965–971 (1995).
61. Gentry, E. C. et al. Reverse metabolomics for the discovery of chemical structures from humans. *Nature* **626**, 419–426 (2024).

62. 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).
63. Mann, A. et al. Palmitoyl serine: An endogenous neuroprotective endocannabinoid-like entity after traumatic brain injury. *J. Neuroimmune Pharmacol.* **10**, 356–363 (2015).
64. Grevengoed, T. J. et al. *N*-acyl taurines are endogenous lipid messengers that improve glucose homeostasis. *Proc. Natl. Acad. Sci.* **116**, 24770–24778 (2019).
65. Waluk, D. P., Vielfort, K., Derakhshan, S., Aro, H. & Hunt, M. C. *N*-Acyl taurines trigger insulin secretion by increasing calcium flux in pancreatic β-cells. *Biochem. Biophys. Res. Commun.* **430**, 54–59 (2013).
66. Aichler, M. et al. *N*-acyl taurines and acylcarnitines cause an imbalance in insulin synthesis and secretion provoking β cell dysfunction in type 2 diabetes. *Cell Metab.* **25**, 1334-1347.e4 (2017).
67. Mannochio-Russo, H. et al. The microbiome diversifies *N*-acyl lipid pools - including short-chain fatty acid derived. In preparation.
68. Inciarte, A. et al. Post-exposure prophylaxis for HIV infection in sexual assault victims. *HIV Med.* **21**, 43–52 (2020).
69. Kamitani, E. et al. Growth in proportion and disparities of HIV PrEP use among key populations identified in the United States national goals: systematic review and meta-analysis of published surveys. *J. Acquir. Immune Defic. Syndr.* **84**, 379 (2020).
70. Walmsley, S. L. et al. Dolutegravir plus abacavir-lamivudine for the treatment of HIV-1 infection. *N. Engl. J. Med.* **369**, 1807–1818 (2013).
71. Clotet, B. et al. Efficacy and safety of darunavir-ritonavir at week 48 in treatment-experienced patients with HIV-1 infection in POWER 1 and 2: a pooled subgroup analysis of data from two randomised trials. *Lancet Lond. Engl.* **369**, 1169–1178 (2007).
72. Huhn, G. D. et al. Darunavir/cobicistat/emtricitabine/tenofovir alafenamide in a rapid-initiation model of care for human immunodeficiency virus type 1 infection: Primary analysis of the DIAMOND study. *Clin. Infect. Dis.* **71**, 3110–3117 (2020).
73. Sax, P. E. et al. Abacavir–lamivudine versus tenofovir–emtricitabine for initial HIV-1 therapy. *N. Engl. J. Med.* **361**, 2230–2240 (2009).
74. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
75. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, 20 (2015).
76. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).
77. Zuffa, S. et al. microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat. Microbiol.* **9**, 336–345 (2024).
78. Bittremieux, W. et al. Universal MS/MS visualization and retrieval with the metabolomics spectrum resolver web service. Preprint at https://doi.org/10.1101/2020.05.09.086066 (2020).
79. Bittremieux, W., Laukens, K., Noble, W. S. & Dorrestein, P. C. Large-scale tandem mass spectrum clustering using fast nearest neighbor searching. *Rapid Commun. Mass Spectrom.* e9153 (2021).

80. Schmid, R. et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* 1–3 (2023).
81. Nothias, L.-F. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
82. McDonald, D. et al. Extreme Dysbiosis of the Microbiome in Critical Illness. *mSphere* **1**, e00199-16 (2016).
83. Brennan, C. et al. Clearing the plate: a strategic approach to mitigate well-to-well contamination in large-scale microbiome studies. *mSystems* **0**, e00985-24 (2024).
84. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).