Metabolite discovery through global annotation of untargeted metabolomics data

Li Chen[1,2], Wenyun Lu[2,3], Lin Wang[2,3], Xi Xing[2,3], Xin Teng[2], Xianfeng Zeng[2,3], Antonio D.Muscarella[2], Yihui Shen[2], Alexis Cowan[2,4], Melanie R. McReynolds[2,3], Brandon Kennedy[5], Ashley M. Lato[6], Shawn R. Campagna[6], Mona Singh[2,7], Joshua Rabinowitz[2,3,4,#]

[1]Institute of Metabolism and Integrative Biology, Fudan University, Shanghai, 200438, China.

[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, 08544, USA.

[3]Department of Chemistry, Princeton University, Princeton, NJ, 08544, USA.

[4]Department of Molecular Biology, Princeton University, Princeton, NJ, 08544, USA.

[5]Lotus Separation LLC, Department of Chemistry, Princeton University, Princeton, NJ, 08544, USA

[6]Department of Chemistry, The University of Tennessee at Knoxville, Knoxville, TN, 37996, USA

[7]Department of Computer Science, Princeton University, Princeton, NJ, 08544, USA.

Corresponding author, e-mail: joshr@princeton.edu

Abstract

A primary goal of metabolomics is to identify all biologically important metabolites. One powerful approach is liquid chromatography-high resolution mass spectrometry (LC-MS), yet most LC-MS peaks remain unidentified. Here, we present a global network optimization approach, NetID, to annotate untargeted LC-MS metabolomics data. We consider all experimentally observed ion peaks together, and assign annotations to all of them simultaneously so as to maximize a score that considers properties of peaks (known masses, retention times, MS/MS fragmentation patterns) as well network constraints that arise based on mass difference between peaks. Global optimization results in accurate peak assignment and trackable peak-peak relationships. Applying this approach to yeast and mouse data, we identify a half-dozen novel metabolites, including thiamine and taurine derivatives. Isotope tracer studies indicate active flux through these metabolites. Thus, NetID applies existing metabolomic knowledge and global optimization to annotate untargeted metabolomics data, revealing novel metabolites.

Introduction

Metabolomics provides a snapshot of small-molecule concentrations in a biological system. In so doing, it reflects the integrated impact of genetics and the environment on metabolism. One important role of metabolomics is annotating previously unknown or underappreciated metabolites. For example, metabolomics facilitated identification of 2-hydroxyglutarate as an oncometabolite, eventually leading to the development of inhibitors of 2-hydroxyglutarate synthesis as anticancer agents[1,2]. Metabolomics also contributed to identification of a diversity of natural products[3,4] and disease biomarkers[5].

A common experimental strategy in metabolomics is liquid chromatography-high resolution mass spectrometry (LC-MS). LC-MS metabolomics measures thousands of ion peaks, of which hundreds are associated with known metabolites. A much greater number of peaks, however, still remain unannotated. The standard approach to peak annotation is to compare exact mass and either retention time or MS/MS fragmentation pattern to authenticated standards. To facilitate such comparisons, extensive chemical databases have been developed (e.g. METLIN[6], HMDB[7], MoNA[8], KEGG[9], Pubchem[10], ChEBI[11] and NIST[12]), with software tools available for automated peak picking and database comparison. Modern software also includes features for annotating peaks arising from isotopes and adducts of known metabolites, based on co-elution and characteristic mass differences (e.g. XCMS[13,14], GNPS[15], MS-DIAL[16], MZmine[17], and CAMERA[18]). Such peaks seem to account for at least half of non-background LC-MS features[19,20]. Despite this progress, a great number of unknown peaks remain, and figuring out their identities is a primary challenge in the field.

One promising approach is network analysis, capitalizing on peak-peak relationships to increase annotation scope and accuracy. Connections can be drawn based on similar responses across experiments and/or MS2 similarity. Such connections can arise either through biochemical activities or mass spectrometry phenomena, such as isotopes, adducts, or in-source fragments. While distinct metabolites typically separate chromatographically, ions connected through mass spectrometry phenomena co-elute. Workflows employing the concept of molecular connectivity have been used to build networks (e.g., GNPS[21,22], CliqueMS[23], MetDNA[24], BioCAn[25], and IPA[26]), and are showing increasing utility for annotating metabolomics data in diverse contexts. For example, GNPS has been used broadly in identifying natural products.

Existing algorithms generally focus on metabolite peaks with MS2 spectra available, using MS2 spectral data as the main annotation driver. This is an effective strategy for annotating high abundance peaks with informative MS2 spectra, such as major secondary metabolites. It is less effective, however, for many low abundance metabolomics peaks, due to poor quality or less informative MS2 spectra. We accordingly set out to develop a network algorithm for annotating the breadth of metabolomics peaks, capitalizing on available MS2 spectra but including also low

abundance peaks lacking MS2 spectra. Effective incorporation of peaks without MS2 spectra required making yet better use of peak-peak relationships to enhance annotation accuracy, which we achieved through the computational approach of global optimization: not dealing with peak annotation one-by-one, but instead all at once to take full advantage of the entire available information. This global optimization strategy had not previously been applied in the context of molecular networking analysis.

To this end, we present the algorithm "NetID". Similar to existing network analysis approaches, nodes are experimentally observed non-background ion peaks and connections are mass differences between peaks. We explicitly distinguish connections due to biotransformations ("biochemical connections" linking two metabolites) from those due to mass spectrometry phenomenon ("abiotic connections" linking isotopes, adducts, and fragments to the metabolites from which they are derived). Peak annotation occurs in a single global optimization step, based on linear programming, that enforces a single formula assignment for each experimentally observed ion peak. Using this approach, we can annotate roughly 80% of untargeted metabolomics peaks, with a majority being isotopes and adducts of known metabolites. Through these efforts, we provide likely formulae for several hundred novel metabolites, and confirm the identities of half-dozen species not currently in metabolomics databases.

Results

NetID algorithm

NetID involves three computational steps: initial annotation, scoring, and optimization (Figure 1). The workflow starts with a peak table that contains a list of peak m/z, RT, intensity, and (when available) associated MS2 spectra, with background peaks removed by comparing to a process blank sample. Each peak defines a node in the network. In the initial annotation phase, we match every experimentally measured node m/z to formulae in the HMDB database. Peaks matching to HMDB formula within 10 ppm are annotated as seed nodes, from which we extend edges to build the network.

Edges connect two nodes via gain or loss of specific chemical moieties (atoms). The atom differences can occur either due to metabolism (biochemical connection) or due to mass spectrometry phenomena (abiotic connections). For example, a difference of H_2 suggests an oxidation/reduction relationship and defines a biochemical edge. A difference of Na-H suggests sodium adducting and is a type of abiotic edge (adduct edge). Other atom differences define other types of abiotic connections (isotope or fragment edges). Most atom differences are specific to biochemical, adduct, isotope, or fragment edges, but a few occur in multiple categories. For example, H_2O loss can be either biochemical (enzymatic dehydration) or abiotic (in-source water loss). By integrating literature and in-house data, we assembled a list of 25 biochemical atom differences and 59 abiotic atom

differences which together define all connections in the network (Supplementary Table 1, 2). Using these lists, starting from the seed nodes, we draw all feasible edges such that (i) Δm/z between the connected nodes matches the atom mass difference and (ii) only co-eluting peaks are connected by abiotic edges. Through the edge extension process, possible formulae are assigned to nodes outside the initial seeds. A few rounds of edge extension suffice to give thorough coverage. Due to finite mass measurement precision, a single node may be assigned multiple contradictory formulae, which are resolved at the optimization step (see Methods).

NetID then scores every node and edge annotation. Node annotations are scored based on precision of m/z match to the molecular formula, precision of retention time match to known metabolite retention time and (when the relevant information is available) quality of MS2 spectra match to database structure. In addition, there is a bonus for matching to formula in HMDB and a penalty for breaking basic chemical rules (Seven Golden Rules for filtering molecular formulae[27]). Biochemical edges receive a positive score for MS2 spectra similarity match between the connected nodes, and are otherwise unscored. Abiotic edges are scored based on precision of co-elution with the parent metabolite, connection type (adduct, isotope, etc.), and features specific to the connection type, such as expected natural abundance for isotope peaks (see Methods). The overall impact is to assign high scores to annotations that effectively align the experimentally observed ion peaks with prior metabolomics knowledge.

With a score assigned for each potential node and edge annotation, we formulate the global network optimization problem as that of maximizing the network score with linear constraints that each node and edge has a single unique annotation and that these are consistent (e.g. peaks connected by H_2 edge must have formula differing by 2H). Such optimization is readily performed by linear programing with a typical runtime of hours in R on a personal computer, and results in an optimal and consistent network annotation.

Global network optimization

As an example of the utility of global network optimization, where all peaks and connections are simultaneously considered to enhance annotation accuracy, we present an example network containing five peaks (Figure 2A). We first match experimental measurements to the database, annotating node *a* and node *b* as seed nodes adenosine monophosphate (AMP, $C_{10}H_{14}N_5O_7P$) and adenosine ($C_{10}H_{13}N_5O_4$), respectively. We also identify five possible connections between the five nodes. Two alternative networks are generated by extending annotations. In the left network, node *c* is annotated as adenosine HCl adduct ($C_{10}ClH_{14}N_5O_4$), whereas in the right network, node *c* is annotated as a putative metabolite ($C_9H_{14}N_5O_5P$) resulting from CO_2 loss from AMP. Node *d* is ^{13}C isotope of node *c* in both networks. Node *e* is annotated as ^{37}Cl isotope of node *c* in the left network, and is unannotated in the right network because there is no Cl atom in the parent molecule.

The left network has higher total node and edge annotation scores than the right network, and thus is selected by NetID. This selection makes sense to an experienced mass spectroscopist: the ^{37}Cl isotope signature in node *e* indicates that node *c* should contain Cl. The power of NetID is that it automatically captures such logic, and uses the power of global computational optimization to extend such inferences across the network in an automated manner.

To test the NetID workflow, we applied it to both yeast and liver datasets, in both positive and negative ionization mode (Figure 2B, 2C). Considering the example of negative mode yeast data with a total of 5,588 non-background peaks, in the initial annotation step, roughly 1,600 potential formulae were assigned to 1,400 peaks, with about 200 peaks receiving multiple formula annotations. These nodes were connected by just over 50,000 potential edges. Edge extension expanded coverage to over 5,000 nodes with an average of twelve potential formulae each, highlighting the importance of scoring and network optimization to assign proper formulae. After scoring node and edge annotations, global network optimization settled on about 4,800 unique node annotations. About 20% of the annotated peaks were metabolites, 14% were putative novel metabolites, and the rest were mass spectrometry phenomena, such as adducts, fragments, isotopes. Nodes were connected by about 10,000 edges, roughly evenly split between biochemical and abiotic connections (Figure 2C, Supplementary Fig. 1A). More than 90% of annotated nodes fell into a single dominant connected network (Supplementary Fig. 1B), reflecting most peaks being connected to core metabolism. About 15% of peaks, however, remained unannotated. These unannotated peaks likely reflect deficiencies in our lists of allowed atom differences, including additional forms of mass spectrometry phenomena. For example, manual examination of the unconnected peaks revealed a dozen nickel adducts of known compounds (Supplementary Table. 3). Importantly, the annotated peaks included several hundred novel metabolite formulae (Supplementary Fig. 2, Supplementary Data 1). Collectively, these provide a wealth of opportunities for metabolite discovery.

Thiamine-derived metabolites

NetID optimization provided not only a list of putative metabolites, but also connections linking these putative metabolites to known metabolites. In the yeast metabolomics dataset, we found three putative metabolites that have total ion current > 10^5, connected in a subnetwork around thiamine. Their formulae are $C_{12}H_{16}N_4O_2S$ (thiamine+O), $C_{14}H_{20}N_4O_2S$ (thiamine+C_2H_2O) and $C_{14}H_{18}N_4O_2S$, (thiamine+C_2H_4O) (Figure 3A). While not found in HMDB, thiamine+O is documented in METLIN as a thiamine oxidation product, so we focused on the other two potential thiamine derivatives.

MS/MS spectra of the putative thiamine+C_2H_2O and thiamine+C_2H_4O contained characteristic thiamine fragments. Both contained a classical pyrimidine fragment, with thiamine+C_2H_2O also containing an acetylated pyrimidine fragment, leading to a probable structure (Figure 3A,B). The structural assignment is further supported by the presence of an unmodified thiazole fragment. In

contrast, thiamine+C_2H_4O lacked a classical unmodified thiazole fragment, instead showing a thiazole+C_2H_4O fragment (and a fragment with further water loss) (Figure 3A,B).

Isotope tracing experiments further confirm these two peaks contain thiamine. When fed [U-^{13}C]glucose as sole carbon source, yeast synthesize thiamine *de novo*, resulting fully labeled thiamine species, with carbon counts matching the NetID formula assignments (Figure 3C). Adding unlabeled thiamine to the [U-^{13}C]glucose culture media, yeast uptake the unlabeled thiamine, resulting in unlabeled thiamine and M+2 labeled thiamine+C_2H_2O and thiamine+C_2H_4O species. Although discovered in yeast, these are conserved metabolites, found also in mammalian samples (Figure 3D).

Acetylation is one of the 25 biochemical atom transformations allowed in NetID. The addition of C_2H_4O is much less common biochemically, and was captured in NetID as two steps, acetylation followed by reduction. Accordingly, we looked into thiamine metabolism to explore how thiamine+C_2H_4O might be produced. Thiamine pyrophosphate is an important cofactor in pyruvate dehydrogenase (PDH, the entry step to TCA cycle) (Figure 3E). The de-pyrophosphorylation product of thiamine intermediate in PDH reaction yields thiamine+C_2H_4O matches the proposed thiamine+C_2H_4O structure (Figure 3F).

Based on this biochemical route, we realized that analogous products could be formed by α-ketoglutarate dehydrogenase (thiamine+$C_4H_6O_3$) and branched-chain keto acid dehydrogenase (thiamine+C_4H_8O) (Figure 3F). Peaks at both of these exact masses were also experimentally observed, with isotope labeling results supporting their being thiamine-derived metabolites (Supplementary Fig. 3). Thus, NetID enabled the discovery of four novel thiamine-derived metabolites.

N-glucosyl-taurine

We similarly carried out NetID annotation of a mouse liver dataset. We observed multiple putative metabolite peaks linked to taurine, by apparent glucosylation (+$C_6H_{10}O_5$), palmitylation (+$C_{16}H_{30}O$) and transamination (+O-NH_3) (Figure 4A). The latter two, while missing in HMDB, were found in METLIN: N-palmitoyl taurine ($C_{18}H_{37}NO_4S$) and sulfoacetaldehyde ($C_2H_4O_4S$). To elucidate the structure of the putative taurine glucosylation product ($C_8H_{17}NO_8S$), we chemically synthesized N-glucosyl-taurine. Synthetic N-glucosyl-taurine matched the retention time and MS/MS fragmentation pattern of the observed $C_8H_{17}NO_8S$ peak (Figure 4B,C). In liver samples of mice infused with [U-^{13}C]glucose, $C_8H_{17}NO_8S$ appeared in M+6 form, suggesting active synthesis of the N-glucosyl-taurine from circulating glucose (Figure 4D). N-glucosyl-taurine was not observed in yeast extract but was detected in multiple mouse tissues. Quantitation using the synthetic standard shows that liver has the highest level of glucosyl-taurine at ~170 µM (Figure 4E, Supplementary Fig. 4). This ranks among the few dozen most abundant liver metabolites.

Discussion

The advent of LC-MS metabolomics revealed tens of thousands of metabolite peaks not matching known formulae, raising the possibility that the majority of metabolites remained to be discovered. While the biosphere likely contains many novel metabolites, it has been increasingly recognized that most peaks in typical untargeted metabolomics studies do not arise from novel metabolites, but rather mass spectrometry phenomena. The goal of comprehensively annotating untargeted metabolomics peaks with molecular formulae has, however, remained elusive.

One promising strategy for peak annotation involves building molecular networks where nodes are LC-MS peaks (with associated molecular formulae) and edges are atom transformations linking the peaks. Here we advance this strategy by combining metabolomics knowledge with computational global optimization. We explicitly differentiate biochemical connections reflecting metabolic activity and abiotic connections arising from mass spectrometry phenomena. By formulating the peak annotation challenge as a linear program, we identify an optimal network in light of all observed peaks. Rather than weeding out peaks from mass spectrometry phenomena like adducts and natural isotopes, this approach takes advantage of the information embedded in them. It further provides traceable peak-peak relationships, which illuminate the basis for assigned formulae and suggest candidate structures.

Applying this approach to untargeted LC-MS data from yeast and liver samples, we assign formulae to roughly three-quarters of all non-background peaks. In each of positive and negative mode, the annotated peaks cover about 1000 known metabolites, with on average more than four mass peaks for every metabolite (e.g. M+H plus three adduct or isotope peaks). This leaves a couple thousand unannotated peaks from each LC-MS run. Based on the observed ratio between peaks and metabolites, this likely correspond to hundreds (but not thousands) of unidentified metabolites. This number may actually be less, due to novel adducts (e.g. nickel adducts, which we discovered via careful examination of the unannotated peaks) or other mass spectrometry phenomena. Importantly, this approach has already generated likely formulae for many hundreds of putative novel metabolites (Supplementary Fig. 2, Supplementary Data 1), including a half-dozen for which we assign structures (Figure 3, 4).

A key benefit of molecular network-based annotation is the ability to assimilate steadily new information[21,22]. Each newly identified metabolite provides an additional anchor point for optimizing the network. Other data types can be seamlessly added. For example, compound class categorization based on MS/MS data[28] or retention time prediction[29] can be added to score nodes. Labeling similarity upon feeding different isotope-labeled nutrients could potentially be added to score edges. Global optimization, integrating all new information comprehensively with prior knowledge to arrive at optimal annotations, is novel and potentially transformative for the field more broadly. The cycle

of careful experimentation and focused computational method developments holds the potential to identify most unknown metabolites over the coming decade, providing a robust blueprint of the metabolome (Figure 5).

Methods

Yeast metabolomics sample preparation and isotope labeling

S. cerevisiae strain FY4 was grown for at least 10 generations in minimal essential media containing 0.4% [U-^{12}C] or [U-^{13}C] glucose and 10 mM ammonium sulfate with or without 0.4 mg/L thiamine hydrochloride[30]. Then, in mid-exponential phase, 5 mL culture broth (OD_{600} = 0.80) was filtered and metabolites were extracted using 1 mL extraction buffer (40:40:20:0.5 acetonitrile:methanol:water:formic acid), followed by adding 88 μL neutralization buffer (15% NH_4HCO_3). The extracts were kept at -20℃ for at least15 min to precipitate protein before centrifuging at 16,000 g for 10 min. The supernatant was used for LC–MS analysis.

Murine metabolomics sample preparation and intravenous infusion experiment

Animal studies followed protocols approved by the Princeton University Institutional Animal Care and Use Committee. Twelve-month-old female wild-type C57BL/6 mice (The Jackson Laboratory, Bar Harbor, ME) on normal diet were sacrificed by cervical dislocation and tissues quickly dissected and snap frozen in liquid nitrogen with precooled Wollenberger clamp. Frozen samples from liquid nitrogen were then transferred to −80°C freezer for storage. To extract metabolites, frozen liver tissue samples were first weighed (~ 20 mg each) and transferred to 2 mL round-bottom Eppendorf Safe-Lock tubes on dry ice. Samples were then ground into powder with a cryomill machine (Retsch, Newtown, PA) for 30 seconds at 25 Hz, and maintained at cold temperature using liquid nitrogen. For every 25 mg tissues, 922 uL extraction buffer (as above) was added to the tube, vortexed for 10 seconds, and allowed to sit on ice for 10 minutes. Then 78 μL neutralization buffer was added and the samples vortexed. The samples were allowed to sit on ice for 20 minutes and then centrifuged at 16,000 g for 25 min at 4°C. The supernatants were transferred to another Eppendorf tube and centrifuged at 16,000 g for another 25 min at 4°C. The supernatants were transferred to glass vials for LC-MS analysis. A procedure blank was generated identically without tissue, which was used later to remove the background ions.

Detailed methods for intravenous infusion of mice have been described previously[31]. Briefly, *in vivo* infusions were performed on 12–14-week-old C57BL/6 mice pre-catheterized in the right jugular vein (Charles River Laboratories). Mice were kept fasted for 6 h and then infused for 2.5 h with [U-^{13}C]glucose (200 mM, 0.1 μL/min/g). The mouse infusion setup (Instech Laboratories) included a tether and swivel system so that the animal had free movement in the cage. Venous samples were taken from tail bleeds. At the end of the infusion, the mouse was euthanized by cervical dislocation and tissues were collected and extracted as above. Serum metabolites were extracted by adding 100 μl methanol to 5 μL of serum and centrifuging for 20 min. The supernatant was used for LC–MS analysis.

LC-MS and LC-MS/MS

LC separation was achieved using a Vanquish UHPLC system (Thermo Fisher Scientific) with an Xbridge BEH Amide column (150×2mm, 2.5 µm particle size; Waters). Solvent A is 95:5 water: acetonitrile with 20 mM ammonium acetate and 20 mM ammonium hydroxide at pH 9.4, and solvent B is acetonitrile. The gradient is 0 min, 90% B; 2 min, 90% B; 3 min, 75%; 7 min, 75% B; 8 min, 70%, 9 min, 70% B; 10 min, 50% B; 12 min, 50% B; 13 min, 25% B; 14 min, 25% B; 16 min, 0% B, 20.5 min, 0% B; 21 min, 90% B; 25 min, 90% B. Total running time is 25 min at a flow rate of 150 µl/min. LC-MS data were collected on a Q-Exactive Plus mass spectrometer (Thermo Fisher) operating in full scan mode with a MS1 scan range of m/z 70-1000, and resolving power of 160,000 at m/z 200. Other MS parameters are as follows: sheath gas flow rate, 28 (arbitrary units); aux gas flow rate, 10 (arbitrary units); sweep gas flow rate, 1 (arbitrary units); spray voltage, 3.3 kV; capillary temperature, 320°C; S-lens RF level, 65; AGC target, 3E6 and maximum injection time, 500 ms.

To demonstrate the utility of inclusion of MS2 data for NetID analysis, 1479 and 803 MS2 spectra were obtained for selected peaks with intensity > 10^5 in positive and negative ionization mode respectively from a previous liver dataset[32]. Targeted MS2 spectra were collected using the PRM function at 25 eV HCD energy with other instrument setting being, resolution 17500, AGC target 10^6, Maximum IT 250 ms, isolation window 1.5 m/z.

Glucosyl-taurine synthesis

Glucosyl-taurine synthesis was carried out following previous literature reports with slight modifications[33]. In brief, dry methanol was obtained by distillation of HPLC-grade methanol (Fisher; HPLC grade 0.2 micron filtered) over CaH_2 (Acros Organics; ca. 93% extra pure, 0-2 mm grain size). A flame-dried round-bottom flask equipped with a reflux condenser and stir bar was charged with 2.0 g taurine (Alfa Aesar; 99%), 3.1 g D-glucose (Acros Organics; ACS reagent), and 80 mL of dry methanol. This mixture was sonicated under an inert atmosphere for 30 minutes before being returned to the manifold for the reaction. To the fine-suspension of taurine and glucose in dry methanol at room temperature, 4.0 mL 5.4 M sodium methoxide in methanol (Acros Organics) was added via glass syringe. At this point, the suspension began to dissolve and after 30 minutes, gave a clear and colorless solution. The solution was stirred vigorously under an inert atmosphere for 72 hours, which resulted in a faint peach-colored solution. This solution was chilled to 0 ˚C, and ~200 mL of absolute ethanol (200 proof) was added and precipitation was allowed to occur at this temperature for 30 minutes. Solvent was then removed by filtration over a glass filter (medium porosity), and washed with ~100 mL of absolute ethanol, affording a fine pale-yellow powder (2.4 g; crude material).

NMR experiment was carried out to validate the structure of synthesized N-glucosyl-taurine. Selective TOCSY experiments using DIPSI2 spin-lock and with added chemical shift filter[34] were run on a Bruker Avance III HD NMR spectrometer equipped with a custom-made QCI-F cryoprobe (Bruker, Billerica, MA) at 800 MHz and at 295.2K controlled temperature. The sample was dissolved in DMSO-

d6. The spectra shown on the plots are results of 200 ms SL mixing, 8 scans each. Data processing (MNova v.14, Mestrelab Research S.L., Santiago de Compostela, Spain) included zero filling, 1 Hz Gaussian apodization, phase- and baseline correction. NMR analysis suggests that the final crude material contains 5.2% N-glucosyl-taurine and unreacted substrates (Supplementary Figure 5).

NetID algorithm

I. Data preparation and input

LC-MS raw data files (.raw) were converted to mzXML format using ProteoWizard[35] (version 3.0.11392). El-MAVEN (version 7.0) was used to generate a peak table containing m/z, retention time, intensity for peaks. Parameters for peak picking were the defaults except for the following: mass domain resolution is 10 ppm; time domain resolution is 15 scans; minimum intensity is 1000; minimum peak width is 5 scans. The resulting peak table was exported to a .csv file. Redundant peak entries due to imperfect peak picking process are removed if two peaks are within 0.1 min and their m/z difference are within 2 ppm. Background peaks are removed if its intensity in procedure blank sample is > 0.5-fold of that in biological sample.

The m/z of the remaining peaks are recalibrated by applying an absolute m/z adjustment factor $\varepsilon_{absolute}$ (independent of measured m/z) and a relative m/z adjustment factor $\varepsilon_{relative}$ (linearly dependent on measured m/z). For each peak *i* the recalibrated values $i_{m/z,\ adjusted}$ are computed as

$$i_{\mathrm{m/z,adjusted}} = i_{\mathrm{m/z,measured}} \times (1 + \varepsilon_{\mathrm{relative}}) + \varepsilon_{\mathrm{absolute}} \quad (1)$$

The $\varepsilon_{relative}$ and $\varepsilon_{absolute}$ values are fit via linear regression using measured m/z values of selected known metabolite ion peaks and their calculated m/z. That is, for each of these known metabolite *k*, we have equations

$$k_{\mathrm{m/z,calculated}} = k_{\mathrm{m/z,measured}} \times (1 + \varepsilon_{\mathrm{relative}}) + \varepsilon_{\mathrm{absolute}} \quad (2)$$

LC-MS/MS data were extracted from the mzXML files using lab-developed Matlab code. MS2 spectra may contain interfering product ions from co-eluting isobaric parent ions. These interfering product ions were removed by examining the extracted ion chromatogram (EIC) similarity between the product ions in MS2 data and the parent ion in MS1 data. A Pearson correlation coefficient of 0.8 was used as a cutoff to retain those product ions that has similar EIC as the parent ion. The cleaned MS2 data were exported to Excel files for further processing.

Structures, formulae, m/z and MS2 spectra of metabolites were obtained from the Human Metabolome Database (HMDB, version 4.0), and retention times of selected metabolites were determined through running authentic standards using the above-mentioned LC-MS method.

NetID algorithm requires three types of input files: a peak table (in .csv format) recording m/z, retention time, intensity for peaks; an atom difference rule table (in .csv format) containing a list of 25 biochemical atom differences and 59 abiotic atom differences which together define all connections in the network (Supplementary Table 1, 2), and metabolite information files containing structure, formula, m/z and MS2 spectra of HMDB metabolites and retention time of selected metabolites under different LC conditions. Exemplary peak table from the yeast dataset, atom difference rule table and HMDB metabolite information file are provided in Supplementary Data 2.

II. Initial annotation of nodes and edges in the network

The first step of NetID algorithm is to make an initial annotation for seed nodes, determine possible annotations for other nodes, and determine edges in the network. Each peak is a node in the network. We compare the experimentally measured m/z for each node to those of all metabolite formulae in the HMDB database. When the m/z difference is within 10 ppm, candidate formulae and HMDB IDs are assigned to the node, and this node is defined as a primary seed node. A primary seed node can contain more than one candidate formulae and HMDB IDs if all are within the m/z difference range.

Edges connect two nodes via gain or loss of specific atoms. We assembled a list of 25 biochemical atom differences and 59 abiotic atom differences which together define all connections in the network (Supplementary Table 1, 2). Let each of these differences be denoted by D_i. For each node u, if there is a node v such that the difference in the measured m/z of the nodes matches one of the those in the list of atom mass differences, we add an edge between u and v. That is, if $u_{\mathrm{m/z}}$ and $v_{\mathrm{m/z}}$ are the experimentally measured m/z for the peaks corresponding to nodes u and v respectively (assuming $v_{\mathrm{m/z}} > u_{\mathrm{m/z}}$ for simplicity), then there is an edge between these nodes if there is some difference D_i such that

$$\left| \left(v_{\mathrm{m/z}} - u_{\mathrm{m/z}}\right) - D_i \right| < v_{\mathrm{m/z}} \times 10 \text{ ppm} \tag{3}$$

If D_i is an abiotic difference, in order to add an edge, it is additionally required that the retention time between two nodes should be within 0.2 min. That is, if u_{RT} and v_{RT} are the retention times for u and v respectively, then it is required that

$$\left| v_{\mathrm{RT}} - u_{\mathrm{RT}} \right| < 0.2 \text{ min} \tag{4}$$

For each node, its candidate formulae set will expand due to propagating formulae from its neighboring nodes through edge atom differences. For example, when applying the atom difference of edge (u, v) on the formula assigned to primary seed node u, we can derive a new candidate formula for the connected node v. If the derived formula's calculated m/z is within 5 ppm of node v's measured m/z, then a new candidate formula is added for node v. Iterating the process to all candidate formulae of node u through edge (u, v) will further expand candidate formulae for node v.

We apply the above extension process to formulae of all primary seed nodes through atom difference edges, and these new candidate formulae can themselves be used for another round of extension. Note that a primary seed node will be treated as the rest of nodes during the subsequent rounds of extension, and may as well be assigned with new formulae.

To avoid duplicated efforts in the extension process, we allow formulae of primary seed nodes and biotransformed formulae thereof to be extended through both biotransformation and abiotic atom difference edges, and do not allow abiotic candidate formulae be further extended through biotransformation atom difference edges. The default extension process includes two rounds of biotransformation edge extensions and three rounds of abiotic edge extensions.

III. Scoring node annotations

NetID then scores every candidate node and edge annotation assigned in the initial annotation step. The node scoring system aims to assign high scores to annotations that align observed ion peaks with known metabolites based on m/z, retention time, MS/MS, and/or isotope abundances.

Let the set of candidate annotation for node *u* be denoted as $\{a_1 \dots a_i \dots a_m\}$. For each node *u* and each of its candidate annotation a_i, let S(u, a_i) denotes the score of candidate annotation a_i for node *u*. Different scoring components for candidate node annotations are defined as below:

(a) $S_{m/z}(u, a_i)$ is negative when measured m/z differs from the calculated m/z of assigned molecular formula. A larger ppm difference between calculated formula m/z and measurement m/z results to lower scores. The default scale factor is -0.5. Let $a_{i,\mathrm{m/z}}$ be the calculated formula m/z of annotation a_i, and $u_{\mathrm{m/z}}$ be the measured m/z of node *u*, then

$$\mathrm{S_{m/z}}(u, a_i) = -0.5 \times \left| u_{\mathrm{m/z}} - a_{i,\mathrm{m/z}} \right| / u_{\mathrm{m/z}} \times 10^6 \quad (5)$$

(b) $S_{RT}(u, a_i)$ is positive if the measured RT for the peak corresponding to node *u* matches to a known standard. A smaller difference between known and measured RT results in a higher score. Let $a_{i,\mathrm{RT}}$ is the known RT of annotation a_i, and u_{RT} be the measured RT of node *u*, then

$$\mathrm{S_{RT}}(u, a_i) = 1 - \left| u_{\mathrm{RT}} - a_{i,\mathrm{RT}} \right|, \text{ if } \left| u_{\mathrm{RT}} - a_{i,\mathrm{RT}} \right| < 0.5 \text{ min}$$

$$\text{Otherwise, } \mathrm{S_{RT}}(u, a_i) = 0 \quad (6)$$

(c) $S_{MS2}(u, a_i)$ is positive if the measured MS2 spectrum of node *u* matches the database MS2 spectrum of annotation a_i. A dot product scoring function is used to score the MS2 spectra similarity[24]. The intensities of the fragment ions in the MS2 spectra are rescaled so that the highest fragment ion is set to 1. MS2 spectra are represented as W = [relative intensity of MS2 ions]^n[m/z value]^m, with $\mathrm{n} = 1, \mathrm{m} = 0$. Dot product (DP) and score for MS2 match ($S_{MS2}(u, a_i)$) are defined as below.

$$DP = \frac{\sum W_u W_{a_i}}{\sqrt{\sum W_u^2 \times \sum W_{a_i}^2}} \tag{7}$$

$$\mathrm{S_{MS2}}(u, a_i) = \mathrm{DP}, \text{ if } \mathrm{DP} > 0.5$$

$$\text{Otherwise } \mathrm{S_{MS2}}(u, a_i) = 0 \tag{8}$$

(d) $S_{database}(u, a_i)$ is positive if the annotated formula a_i exists in HMDB. We give a positive score to a primary seed node annotation if that annotated formula exists in HMDB.

$$\mathrm{S_{database}}(u, a_i) = 0.5, \text{ if } a_i \text{ in HMDB}$$

$$\text{Otherwise, } \mathrm{S_{database}}(u, a_i) = 0 \tag{9}$$

(e) $S_{missing\_isotope}(u, a_i)$ is negative if an isotopic peak is missing. We penalize a formula annotation if it passes the intensity threshold (default at $5x10^4$) but does not have isotopic peaks of specified elements. The default isotope being evaluated is ^{37}Cl. Any other elements, such as ^{13}C or ^{18}O, can be included by users.

$$\mathrm{S_{missing\_isotope}}(u, a_i) = -1, \text{ if isotopic peak is missing}$$

$$\text{Otherwise } \mathrm{S_{missing\_isotope}}(u, a_i) = 0 \tag{10}$$

(f) $S_{rule}(u, a_i)$ is negative if annotation a_i violates basic chemical rules. We strongly penalize formulae that violate basic chemical rules, including a negative RDBE (ring and double bond equivalents), and unlikely element ratios in metabolites (O/P < 3, O/Si < 2).

$$\mathrm{S_{rule}}(u, a_i) = -10, \text{ if chemical rules are violated}$$

$$\text{Otherwise, } \mathrm{S_{rule}}(u, a_i) = 0 \tag{11}$$

(g) $S_{derivative}(u, a_i)$ is positive if the annotation a_i is derived from a parent peak *p* with an annotation *h* that has high score $S_{parent}(p, h)$, which is calculated by summing up scores in (a)-(f) for $S(p, h)$.

$$\mathrm{S_{derivative}}(u, a_i) = \mathrm{S_{parent}}(p, h) - 0.5 \tag{12}$$

$$\mathrm{S_{parent}}(p, h) = \mathrm{S_{m/z}}(p, h) + \mathrm{S_{RT}}(p, h) + \mathrm{S_{MS2}}(p, h) + \mathrm{S_{database}}(p, h) + \mathrm{S_{missing\_isotope}}(p, h) + \mathrm{S_{rule}}(p, h) \tag{13}$$

This is particularly helpful in annotating abiotic peaks. For example, annotation of glutamate sodium adduct will be given a positive $S_{derivative}$ when its parent node is annotated as glutamate with high S_{parent} score.

A final score $S(u, a_i)$ for each candidate annotation a_i of node *u* is calculated by summing scores in (a)-(g).

$$\mathrm{S}(u, a_i) = \mathrm{S}_{\mathrm{m/z}}(u, a_i) + \mathrm{S}_{\mathrm{RT}}(u, a_i) + \mathrm{S}_{\mathrm{MS2}}(u, a_i) + \mathrm{S}_{\mathrm{database}}(u, a_i) + \mathrm{S}_{\mathrm{missing\_isotope}}(u, a_i) + \mathrm{S}_{\mathrm{rule}}(u, a_i) + \mathrm{S}_{\mathrm{derivative}}(u, a_i) \quad (14)$$

Note that for each node *u*, we have one of candidate "annotations" that corresponds to no annotation being chosen for that node. The node score for this null annotation is 0 at default, and can be set at a negative value to promote choosing actual annotations.

IV. Scoring edge annotations (biological, adduct, isotope)

The edge scoring system aims to assign high scores to edge annotations that correctly capture biochemical connections between metabolites (based on MS2 spectra similarity) and abiotic connections between metabolites and their mass spectrometry phenomena derivatives, such as isotopes and adducts. Biochemical, isotope, and adduct edge annotations are the most common types, and other less common abiotic connection types are then described in the subsequent section.

Suppose we consider two nodes *u* and *v* that are connected by an edge (*u*, *v*). For each pair of nodes *u* and *v* such that there is an edge (*u*, *v*), let the set of candidate formula for node *u* and *v* be denoted as $\{a_1 \dots a_i \dots a_m\}$ and $\{b_1 \dots b_j \dots b_n\}$, respectively, and let the set of candidate atom differences for edge (*u*, *v*) be $\{D_1 \dots D_k \dots D_l\}$. Let $\mathrm{S}(u, v, a_i, b_j, D_k)$ be the score of choosing candidate formula a_i for node *u*, candidate formula b_j for node *v* and candidate atom difference D_k for edge (*u*, *v*). Note that $\mathrm{S}(u, v, a_i, b_j, D_k)$ is set to be 0 if atom difference D_k does not represent the formula difference of a_i and b_j.

$$\mathrm{S}(u, v, a_i, b_j, D_k) = 0, \text{ if } |a_i - b_j| \neq D_k$$

Different scoring components for candidate edge annotations are defined as below:

(h) When node *u* and *v* have experimental measured MS2 spectra, $\mathrm{S}_{\mathrm{MS2\_similarity}}(u, v, a_i, b_j, D_k)$ is defined for a biochemical edge, and is a positive score if two connected nodes *u* and *v* have MS2 similarity, given the formula difference of a_i and b_j matches the atom difference defined by D_k. $\mathrm{S}_{\mathrm{MS2\_similarity}}$ is determined using the dot product (DP), as described in previous section, and reverse dot product (DP_R), which evaluates the neutral ion loss similarity in the MS2 spectra[24]. A reverse MS2 spectrum is represented as R = [relative intensity of MS2 ions]^n[parent m/z – measured m/z value]^m, with $\mathrm{n} = 1, \mathrm{m} = 0$.

$$\mathrm{DP} = \frac{\sum W_u W_v}{\sqrt{\sum W_u^2 \times \sum W_v^2}} \quad (15)$$

$$\mathrm{DP\_R} = \frac{\sum R_u R_v}{\sqrt{\sum R_u^2 \times \sum R_v^2}} \quad (16)$$

$$\mathrm{S}_{\mathrm{MS2\_similarity}}(u, v, a_i, b_j, D_k) = \max(\mathrm{DP}, \mathrm{DP\_R}), \text{ if } \max(\mathrm{DP}, \mathrm{DP\_R}) > 0.3$$

$$\text{Otherwise, } S_{\text{MS2\_similarity}}(u, v, a_i, b_j, D_k) = 0 \quad (17)$$

(i) $S_{\text{co\_elution}}(u, v, a_i, b_j, D_k)$ is defined for an abiotic edge, and is a negative score if the RT of two connected nodes differ more than a threshold (0.05 min), given the formula difference of a_i and b_j matches the atom difference defined by D_k.

$$S_{\text{co\_elution}}(u, v, a_i, b_j, D_k) = -5 \times |u_{\text{RT}} - v_{\text{RT}}|, \text{ if } |u_{\text{RT}} - v_{\text{RT}}| \geq 0.05 \text{ min}$$

$$\text{Otherwise, } S_{\text{co\_elution}}(u, v, a_i, b_j, D_k) = 0 \quad (18)$$

(j) $S_{\text{type}}(u, v, a_i, b_j, D_k)$ is defined for all edges, given the formula difference of a_i and b_j matches the atom difference defined by D_k, and is a non-negative score depending on the connection type of edge, which is defined by D_k, including biotransformation, adduct, isotope and fragment (Supplementary Table 1, 2). The magnitude of scores reflects the empirical confidence in the annotation type when certain atom differences occur, and can be adjusted based on personal use.

$$S_{\text{type}}(u, v, a_i, b_j, D_k) = 0, \text{ if } D_k \in \text{biotransformation}$$

$$S_{\text{type}}(u, v, a_i, b_j, D_k) = 0.5, \text{ if } D_k \in \text{adduct}$$

$$S_{\text{type}}(u, v, a_i, b_j, D_k) = 2, \text{ if } D_k \in \text{isotope}$$

$$S_{\text{type}}(u, v, a_i, b_j, D_k) = 0.3, \text{ if } D_k \in \text{fragment} \quad (19)$$

(k) For each $D_k \in$ isotope, $S_{\text{isotope\_intensity}}(u, v, a_i, b_j, D_k)$ is defined for isotope edge (*u, v*) where b_j is the isotopic derivative of a_i with atom difference of D_k, and is a negative score if the measured isotope peaks deviate from expected natural abundance. The score for an isotope edge depends on how likely the ratio of measured and expected isotopic intensity ($\text{Ratio}_{\text{isotope}}$) is observed in an empirical normal distribution $\text{N}(1, \sigma^2_{\text{isotope}})$. Isotopes of all elements included in the atom difference table are evaluated.

$$\text{Ratio}_{\text{isotope}} = \frac{v_{\text{intensity}} \,/\, u_{\text{intensity}}}{\text{Expected isotopic intensity ratio } (a_i, b_j, D_k)} \quad (20)$$

$$S_{\text{isotope}_{\text{intensity}}}(u, v, a_i, b_j, D_k) = log_{10}\left[\frac{P\left(\mu = \text{Ratio}_{\text{isotope}} \middle| \text{N}(1, \sigma^2_{\text{isotope}})\right)}{P\left(\mu = 1 \middle| \text{N}(1, \sigma^2_{\text{isotope}})\right)}\right] \quad (21)$$

$\sigma_{isotope}$ is empirically defined as below, so that when measured isotope intensity is close to detection limit, a larger $\sigma_{isotope}$ (a widened distribution, which is more tolerant to discrepancy) will be used.

$$\sigma_{\text{isotope}} = 0.2 + 10^{3-log_{10}(v_{\text{intensity}})} \quad (22)$$

A final edge annotation score $S(u, v, a_i, b_j, D_k)$ for choosing candidate formula a_i for node *u*, candidate formula b_j for node *v* and candidate atom difference D_k for edge (*u*, *v*) is calculated by summing scores in (h)-(k), if other less common abiotic connection types are not considered (see next section).

$$S(u, v, a_i, b_j, D_k) = S_{\text{MS2\_similarity}}(u, v, a_i, b_j, D_k) + S_{\text{co\_elution}}(u, v, a_i, b_j, D_k) +$$

$$S_{\text{type}}(u, v, a_i, b_j, D_k) + S_{\text{isotope\_intensity}}(u, v, a_i, b_j, D_k) \quad (23)$$

V. Additional abiotic edge types

LC-MS metabolomics may include additional abiotic relationships. In orbitrap data, these include oligomers, multi-charge species, heterodimers, in-source fragments of known or unknown metabolites[36], and ringing artifact peaks surrounding high intensity ions[20,37]. These relationships were included in NetID as additional edge types, which are evaluated for all m/z pairs within a predefined RT range (0.2 min).

(l) Oligomer and multi-charge species. An oligomer/multi-charge edge is assigned between two nodes *u* and *v*, if their m/z satisfy

$$|v_{\text{m/z}} - \text{n} \times u_{\text{m/z}}| < u_{\text{m/z}} \times 10 \text{ ppm}, \text{n} \in \{\text{positive integers}\} \quad (23)$$

(m) Heterodimer. Heterodimer peak (node *v*) may be observed when one abundant metabolite (node *u*) forms ion cluster with other ion species (node *t*). We examine nodes that have intensity above 10^5, and assign a heterodimer edge between two nodes *u* and *v* if their m/z difference satisfy

$$|(v_{\text{m/z}} - u_{\text{m/z}}) - t_{\text{m/z}}| < u_{\text{m/z}} \times 10 \text{ ppm} \quad (24)$$

(n) In-source fragments. Fragmentation peaks may be observed when one abundant metabolite breaks up into fragments during the ionization process.

Database MS2 of known metabolites can be used to identify known ion fragmentation peaks[36]. If candidate annotation b_j of node *v* is annotated with a HMDB ID associated with database MS2 spectrum, and m/z of node *u* matches to a fragment m/z in b_j's MS2 spectrum, then a database fragment edge will connect such two nodes. That is,

$$u_{\mathrm{m/z}} \in \text{Database MS2 spectrum of candidate annotation } b_j \text{ of node } v \tag{25}$$

Measured MS2 spectra can be used to identify unknown ion fragmentation peaks. If node *v* is associated with a measured MS2 spectrum, and m/z of another node *u* matches to a fragment m/z in the MS2 spectra, then an experiment fragment edge will connect such two nodes. That is,

$$u_{\mathrm{m/z}} \in \text{Measured MS2 spectrum of node } v \tag{26}$$

(o) Ringing artifacts. Ringing peaks are artifact peaks (node *v*) often observed on both sides of the m/z of an intense ion peak (node *u*) in Fourier-transformed MS instrument including orbitrap. We examine nodes that have intensity above 10^6, and assign a ringing artifact edge between two nodes if two nodes satisfy

$$50 \text{ ppm} < | v_{\mathrm{m/z}} - u_{\mathrm{m/z}} | / u_{\mathrm{m/z}} < 1000 \text{ ppm}$$

$$u_{\mathrm{intensity}} / v_{\mathrm{intensity}} > 50 \tag{27}$$

Scoring of these additional abiotic edges follow the same rules described in the "Scoring edge annotations" section with additional S_{type} defined as below.

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) = 0.5, \text{ if } D_k \in \text{oligomer or multi-charge}$$

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) = 0, \text{ if } D_k \in \text{heterodimer}$$

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) = 0.3, \text{ if } D_k \in \text{database MS2 fragment}$$

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) = 1, \text{ if } D_k \in \text{measured MS2 fragment}$$

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) = 2, \text{ if } D_k \in \text{ringing artifacts} \tag{28}$$

A final edge annotation score $\mathrm{S}(u, v, a_i, b_j, D_k)$ for choosing candidate formula a_i for node *u*, candidate formula b_j for node *v* and candidate atom difference D_k for edge (*u*, *v*) is calculated by summing scores in (h)-(o).

$$\mathrm{S}(u, v, a_i, b_j, D_k) = \mathrm{S_{MS2\_similarity}}(u, v, a_i, b_j, D_k) + \mathrm{S_{co\_elution}}(u, v, a_i, b_j, D_k) +$$

$$\mathrm{S_{type}}(u, v, a_i, b_j, D_k) + \mathrm{S_{isotope\_intensity}}(u, v, a_i, b_j, D_k) \tag{29}$$

VI. Global network optimization using linear programing

Using scores assigned for each candidate node and edge annotation, our goal is to find annotations for each node so as to maximize the sum of the scores across the network under the constraints that each node is assigned a single annotation, and that the network annotation is consistent. We use linear programming to solve this optimization problem optimally, as described next.

For each node *u* and each of its candidate formula a_i, we define a node binary decision variable x_{u,a_i} to denote whether candidate formula a_i is selected as the annotation for node *u*. That is,

$$x_{u,a_i} = 1, \text{ if node } u \text{ is annotated with formula } a_i$$

$$\text{Otherwise, } x_{u,a_i} = 0 \tag{28}$$

We define a binary decision variable c_{u,v,a_i,b_j,D_k} to denote whether candidate formulae a_i and b_j are chosen for nodes *u* and *v* , and the candidate atom difference D_k corresponds to the formula difference of candidate formulae a_i and b_j of the connected nodes *u* and *v*. That is,

$$c_{u,v,a_i,b_j,D_k} = 1, \text{ if } a_i \text{ and } b_j \text{ are chosen for nodes } u \text{ and } v \text{ respectively, and } \left|a_i - b_j\right| = D_k$$

$$\text{Otherwise, } c_{u,v,a_i,b_j,D_k} = 0 \tag{29}$$

We constrain the optimization so that each node has a single annotation, and an edge exists and only exist if the atom difference of that edge annotation matches the formula difference of nodes. As a result, the node and edge binary variables should satisfy

$$\sum_i x_{u,a_i} = 1 \tag{30}$$

$$c_{u,v,a_i,b_j,D_k} \leq x_{u,a_i} \ , \ c_{u,v,a_i,b_j,D_k} \leq x_{v,b_j} \tag{31}$$

$$c_{u,v,a_i,b_j,D_k} \geq x_{u,a_i} + x_{v,b_j} - 1 \tag{32}$$

For all variables defined above, we add the constraints that they are either 0 or 1.

With each candidate node and edge annotation being scored, the objective for the optimization is to find values for all variables $x_{u,a}$ and $c_{u,v,a,b,D}$ so as to maximize the sum of all node scores and edge scores in a network while satisfying the constraints.

$$\text{Maximize: } \sum x_{u,a} \times \mathrm{S}(u,a) \ + \ \sum c_{u,v,a,b,D} \times \mathrm{S}(u,v,a,b,D) \tag{32}$$

The optimization result provides a string of binary numbers that denote if a candidate node or edge annotation is selected for the global optimal network. IBM ILOG CPLEX Optimization Studio (Version 12.8.0 or later) is used to solve the linear programing problem. A cplexAPI package for R is used to

call CPLEX optimization function in an R environment. For the yeast datasets and using the above scoring parameters, optimization finishes within an hour on a standard laptop. Depending on the number of peaks in data tables, the entries in the atom difference tables, and the parameters involved in scoring, runtimes during internal testing ranged from minutes to 48 h.

Code availability

NetID was developed mainly in R, and used a mixture of IBM ILOG CPLEX Optimization Studio, Matlab and Python. NetID code is available for non-commercial use in github at https://github.com/LiChenPU/NetID, under the GNU General Public License v3.0. A ShinyR app is provided to visualize the network results from NetID in a local environment, along with a detailed user guide and example files (Supplementary Note 1, Supplementary Data 2).

Acknowledgement

This work was supported by a Department of Energy (DOE) grant (no. DE-SC0012461 to J.D.R.), the Center for Advanced Bioenergy and Bioproducts Innovation (grant no. DE-SC0018420, subcontract to J.D.R.) and NIH grant R50CA211437 to W.L. M.R.M is funded by the Howard Hughes Medical Institute and Burroughs Wellcome Fund via the PDEP and Hanna H. Gray Fellows Programs. We thank Istvan Pelczer at NMR facility of Department of Chemistry, Princeton University for the NMR analysis, and X. Su for scientific discussion and help. The Center for Advanced Bioenergy and Bioproducts Innovation and the Center for Bioenergy Innovation are both U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Department of Energy.

Competing interests

The authors declare no competing interests.

Author contributions

L.C., M.S. and J.D.R. conceived the project. L.C. developed the NetID algorithm. W.L., L.W., X.Z., A.C. M.M. performed experiments on mouse. L.W., W.L. and L.C. performed experiments on yeast. L.C., W.L., L.W. and X. X. analyze LC-MS and LC-MS/MS data. X.T., A.M. and Y.S. contributed to coding development. B.K., A.M.L., and S.R.C. provided chemical synthesis of taurine-related compounds. L.C. and J.D.R. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Figure legends

Figure 1. A global network optimization approach for untargeted metabolomics data annotation (NetID). The input data are LC-MS peaks with m/z, retention times, intensities and optional MS2 spectra. The output is a molecular network with peaks (nodes) assigned unique formulae and connected by edges reflecting atom differences arising either through enzymatic reaction (biochemical connection) or mass spectrometry phenomenon (abiotic connection). Peaks are classified as "metabolite" (M+H or M-H peak of formula found in HMDB), "putative metabolite" (formula not found in HMDB but with biochemical connection to a metabolite), or "artifact" (only abiotic connection to a metabolite). NetID algorithm involves three steps. Initial annotation first matches peaks to HMDB formulae. These seed annotations are then extended through edges to cover most nodes, with the majority of nodes receiving multiple formula annotations. Each node and edge annotation are then scored based on match to known masses, retention times, and MS/MS fragmentation patterns. Global network optimization maximizes sum of node scores and edge scores, while enforcing a unique formula for each node and unique transformation relationship for each edge.

Figure 2. Utility of global network optimization. (A) An example network demonstrating the value of the global optimization step in NetID. Node *a* and node *b* match HMDB formulae and are connected by an edge of phosphate (HPO_3). Node *c* can be connected to either node *a* or node *b* through mutually incompatible annotations, resulting in two different candidate networks. The table below the two candidate networks shows the annotations and scoring criteria for each, with the left network preferred for more good node and edge annotations. (B) Visualization of the optimal network obtained from negative mode LC-MS analysis of Baker's yeast, containing 4851 nodes and 9699 connections. Metabolite and putative metabolite peaks are in green and artifact peaks in purple. (C) Summary table of NetID annotations of negative and positive mode LC-MS data from Baker's yeast and mouse liver.

Figure 3. NetID reveals thiamine-derived metabolites in yeast. (A) Subnetwork surrounding thiamine. Nodes, connections, and formulae are direct output of NetID. Boxes with structures were manually added. (B) MS2 spectra of thiamine, thiamine+C_2H_2O, and thiamine+C_2H_4O, with proposed structures of the major fragments. (C) Labeling fraction of thiamine and its derivatives, in [U-^{13}C]glucose with and without unlabeled thiamine in the medium. (D) The thiamine derivatives are also found in mouse tissues and urine. (E) Proposed mechanism for formation of thiamine+C_2H_4O. Pyruvate dehydrogenase (PDH) decarboxylates pyruvate, and adds the resulting [C_2H_4O] unit (in red) to thiamine. (F) The same enzymatic mechanism occurs in oxoglutarate dehydrogenase (OGDH) and branched-chain α-ketoacid dehydrogenase complex (BCKDC), and generates thiamine+$C_4H_6O_3$ and thiamine+C_4H_8O respectively.

Figure 4. NetID discovers mammalian taurine derivatives. (A) Subnetwork surrounding taurine from mouse liver extract data. Nodes, connections, and formulae are direct output of NetID. Boxes with structures were manually added. (B) LC-MS chromatogram of N-glucosyl-taurine standard and the putative glucosyl-taurine from liver extract. (C) MS2 spectrum of glucosyl-taurine peak from liver extract (top), and synthetic N-glucosyl-taurine standard (bottom). (D) Isotope labeling pattern of putative glucosyl-taurine in mice, infused via jugular vein catheter for 2 h with [U-^{13}C]glucose. (E) Absolute N-glucosyl-taurine concentration in murine serum and tissues.

Figure 5. NetID applies global optimization for metabolomics data annotation and metabolite discovery.

Reference

1. DiNardo, C. D. *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N. Engl. J. Med.* **378**, 2386–2398 (2018).
2. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739 (2009).
3. Doroghazi, J. R. *et al.* A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology* **10**, 963–968 (2014).
4. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols* **15**, 1954–1991 (2020).
5. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology* **17**, 451–459 (2016).
6. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
7. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46**, D608–D617 (2018).
8. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.* **88**, 7946–7958 (2016).
9. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462 (2016).
10. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **47**, D1102–D1109 (2019).

11. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**, D1214–D1219 (2016).

12. sherena.johnson@nist.gov. NIST Standard Reference Database 1A. *NIST* https://www.nist.gov/srd/nist-standard-reference-database-1a (2014).

13. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **84**, 5035–5039 (2012).

14. Forsberg, E. M. *et al.* Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature Protocols* **13**, 633–651 (2018).

15. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34**, 828–837 (2016).

16. Tsugawa, H. *et al.* A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods* **16**, 295 (2019).

17. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).

18. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).

19. Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.* **142**, 9097–9105 (2020).

20. Wang, L. *et al.* Peak Annotation and Verification Engine for Untargeted LC–MS Metabolomics. *Anal. Chem.* **91**, 1838–1846 (2019).

21. Schmid, R. *et al. Ion Identity Molecular Networking in the GNPS Environment*. http://biorxiv.org/lookup/doi/10.1101/2020.05.11.088948 (2020) doi:10.1101/2020.05.11.088948.

22. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* **17**, 905–908 (2020).

23. Senan, O. *et al.* CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. 8.

24. Shen, X. *et al.* Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nature Communications* **10**, 1516 (2019).

25. Alden, N. *et al.* Biologically Consistent Annotation of Metabolomics Data. *Anal. Chem.* **89**, 13097–13104 (2017).

26. Del Carratore, F. *et al.* Integrated Probabilistic Annotation: A Bayesian-Based Annotation Method for Metabolomic Profiles Integrating Biochemical Connections, Isotope Patterns, and Adduct Relationships. *Anal. Chem.* (2019) doi:10.1021/acs.analchem.9b02354.

27. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**, 105 (2007).

28. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0740-8.

29. Bonini, P., Kind, T., Tsugawa, H., Barupal, D. K. & Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal. Chem.* **92**, 7515–7522 (2020).

30. Xu, Y.-F. *et al.* Discovery and Functional Characterization of a Yeast Sugar Alcohol Phosphatase. *ACS Chem. Biol.* **13**, 3011–3020 (2018).

31. Hui, S. *et al.* Glucose feeds the TCA cycle via circulating lactate. *Nature* **551**, 115–118 (2017).

32. Lu, W. *et al.* Improved Annotation of Untargeted Metabolomics Data through Buffer Modifications That Shift Adduct Mass and Intensity. *Anal. Chem.* **92**, 11573–11581 (2020).

33. Cho, H. J., You, J. S., Chang, K. J., Kim, K. S. & Kim, S. H. Anti-adipogenic Effect of Taurine-Carbohydrate Derivatives. *Bulletin of the Korean Chemical Society* **35**, 1863–1866 (2014).

34. Robinson, P. T., Pham, T. N. & Uhrín, D. In phase selective excitation of overlapping multiplets by gradient-enhanced chemical shift selective filters. *Journal of Magnetic Resonance* **170**, 97–103 (2004).

35. Chambers, M. C. *et al.* A Cross-platform Toolkit for Mass Spectrometry and Proteomics. *Nat Biotechnol* **30**, 918–920 (2012).

36. Xue, J. *et al.* Enhanced in-Source Fragmentation Annotation Enables Novel Data Independent Acquisition and Autonomous METLIN Molecular Identification. *Anal. Chem.* **92**, 6051–6059 (2020).

37. Mitchell, J. M. *et al.* New methods to identify high peak density artifacts in Fourier transform mass spectra and to mitigate their effects on high-throughput metabolomic data analysis. *Metabolomics* **14**, 125 (2018).

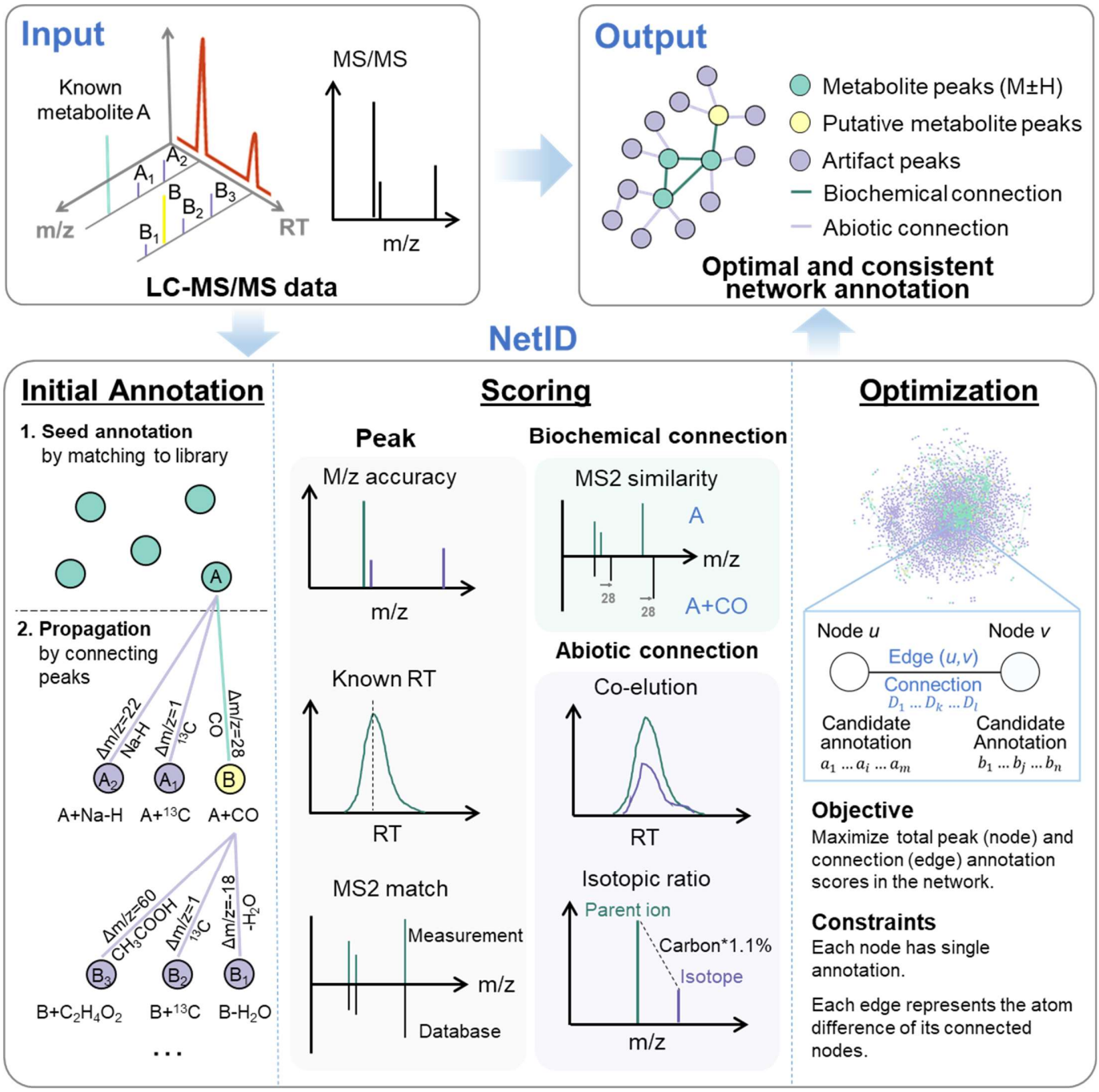
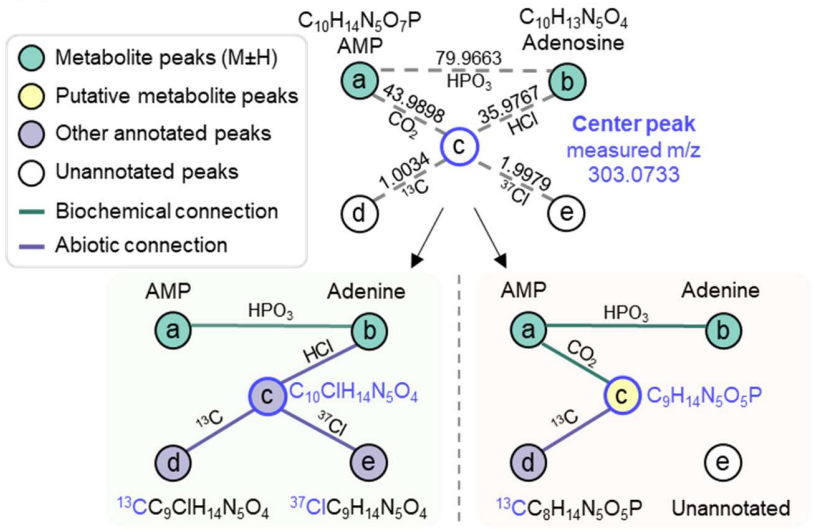


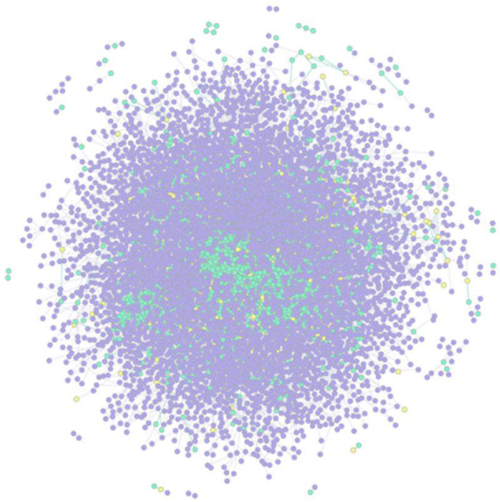
Figure 1. A global network optimization approach for untargeted metabolomics data annotation (NetID). The input data are LC-MS peaks with m/z, retention times, intensities and optional MS2 spectra. The output is a molecular network with peaks (nodes) assigned unique formulae and connected by edges reflecting atom differences arising either through enzymatic reaction (biochemical connection) or mass spectrometry phenomenon (abiotic connection). Peaks are classified as "metabolite" (M+H or M-H peak of formula found in HMDB), "putative metabolite" (formula not found in HMDB but with biochemical connection to a metabolite), or "artifact" (only abiotic connection to a metabolite). NetID algorithm involves three steps. Initial annotation first matches peaks to HMDB formulae. These seed annotations are then extended through edges to cover most nodes, with the majority of nodes receiving multiple formula annotations. Each node and edge annotation are then scored based on match to known masses, retention times, and MS/MS fragmentation patterns. Global network optimization maximizes sum of node scores and edge scores, while enforcing a unique formula for each node and unique transformation relationship for each edge.

A



Node	Annotation	Scoring criteria	Annotation	Scoring criteria
a	$C_{10}H_{14}N_5O_7P$	m/z, RT, MS2	$C_{10}H_{14}N_5O_7P$	m/z, RT, MS2
b	$C_{10}H_{13}N_5O_4$	m/z, RT, MS2	$C_{10}H_{13}N_5O_4$	m/z, RT, MS2
c	$C_{10}ClH_{14}N_5O_4$	m/z	$C_9H_{14}N_5O_5P$	m/z
d	$^{13}CC_9ClH_{14}N_5O_4$	m/z	$^{13}CC_8H_{14}N_5O_5P$	m/z
e	$^{37}ClC_9H_{14}N_5O_4$	m/z	None	
Summary	5 good node scores		4 good node scores	
Edge				
a-b	HPO_3	MS2 similarity	HPO_3	MS2 similarity
a-c	None		CO_2	MS2 similarity
b-c	HCl	Peak shape	None	
c-d	^{13}C isotope	Peak shape & Isotopic ratio	^{13}C isotope	Peak shape & Isotopic ratio
c-e	^{37}Cl isotope	Peak shape & Isotopic ratio	None	
Summary	4 good edge scores		3 good edge scores	
Network	**Total score**	**Optimal network**	**Total score**	**Optimal network**
	Higher	√	Lower	×

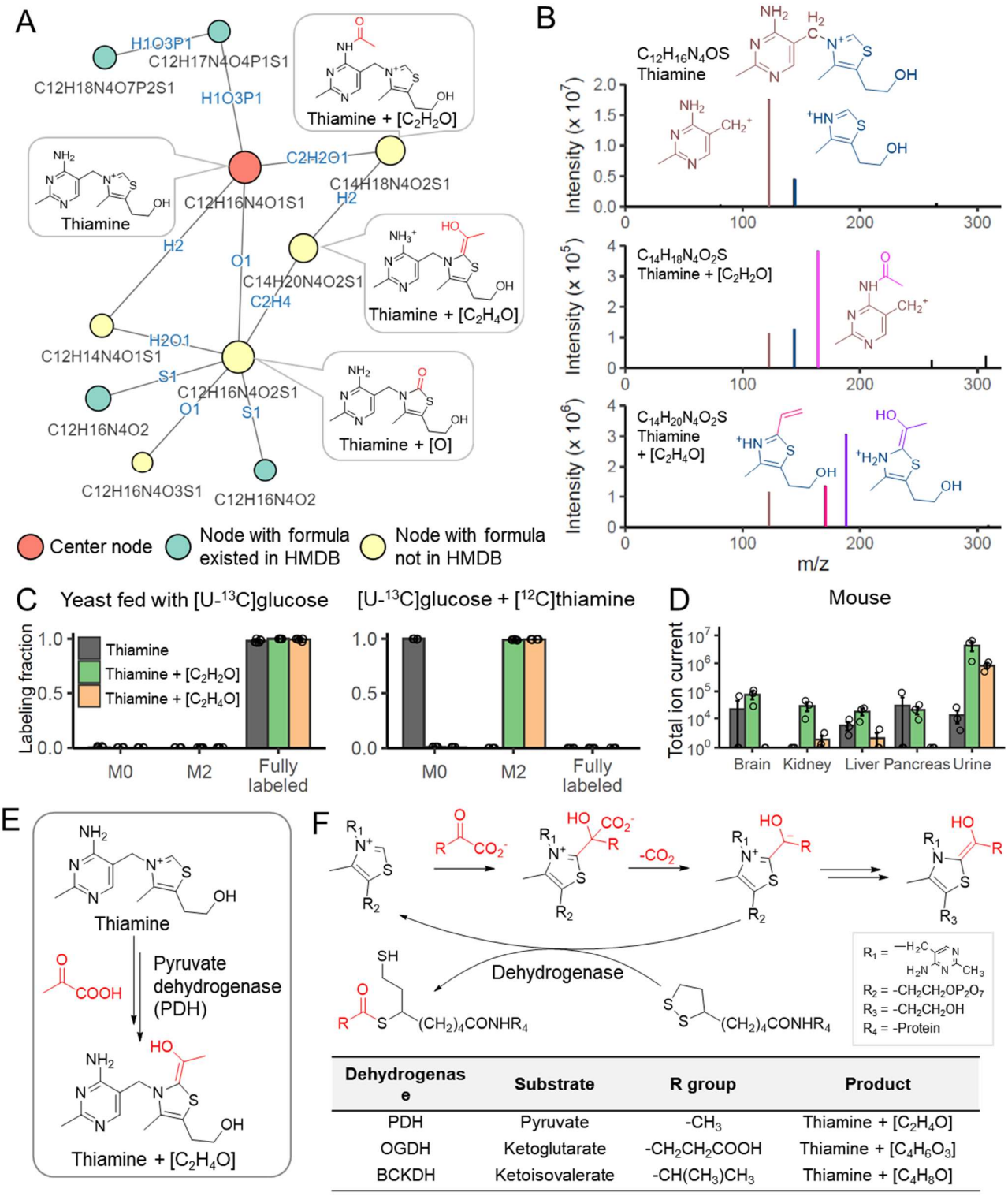
B



C

	Yeast (Neg)	Yeast (Pos)	Liver (Neg)	Liver (Pos)
Total non-background peaks (nodes)	**5588**	**9833**	**8191**	**12128**
Known metabolite	935	1138	1416	1636
Putative metabolite	689	1950	1130	2291
Artifact	3227	5392	4569	6753
Unannotated	737	1353	1076	1448
Total connections (edges)	**9699**	**17007**	**18749**	**18900**
Biochemical	5410	10143	8815	8658
Abiotic	4289	6864	9934	10242

Figure 2. Utility of global network optimization. (A) An example network demonstrating the value of the global optimization step in NetID. Node *a* and node *b* match HMDB formulae and are connected by an edge of phosphate (HPO_3). Node *c* can be connected to either node *a* or node *b* through mutually incompatible annotations, resulting in two different candidate networks. The table below the two candidate networks shows the annotations and scoring criteria for each, with the left network preferred for more good node and edge annotations. (B) Visualization of the optimal network obtained from negative mode LC-MS analysis of Baker's yeast, containing 4851 nodes and 9699 connections. Metabolite and putative metabolite peaks are in green and artifact peaks in purple. (C) Summary table of NetID annotations of negative and positive mode LC-MS data from Baker's yeast and mouse liver.



Dehydrogenase	Substrate	R group	Product
PDH	Pyruvate	$-CH_3$	Thiamine + $[C_2H_4O]$
OGDH	Ketoglutarate	$-CH_2CH_2COOH$	Thiamine + $[C_4H_6O_3]$
BCKDH	Ketoisovalerate	$-CH(CH_3)CH_3$	Thiamine + $[C_4H_8O]$

Figure 3. NetID reveals thiamine-derived metabolites in yeast. (A) Subnetwork surrounding thiamine. Nodes, connections, and formulae are direct output of NetID. Boxes with structures were manually added. (B) MS2 spectra of thiamine, thiamine+C_2H_2O, and thiamine+C_2H_4O, with proposed structures of the major fragments. (C) Labeling fraction of thiamine and its derivatives, in [U-^{13}C]glucose with and without unlabeled thiamine in the medium. (D) The thiamine derivatives are also found in mouse tissues and urine. (E) Proposed mechanism for formation of thiamine+C_2H_4O. Pyruvate dehydrogenase (PDH) decarboxylates pyruvate, and adds the resulting [C_2H_4O] unit (in red) to thiamine. (F) The same enzymatic mechanism occurs in oxoglutarate dehydrogenase (OGDH) and branched-chain α-ketoacid dehydrogenase complex (BCKDC), and generates thiamine+$C_4H_6O_3$ and thiamine+C_4H_8O respectively.

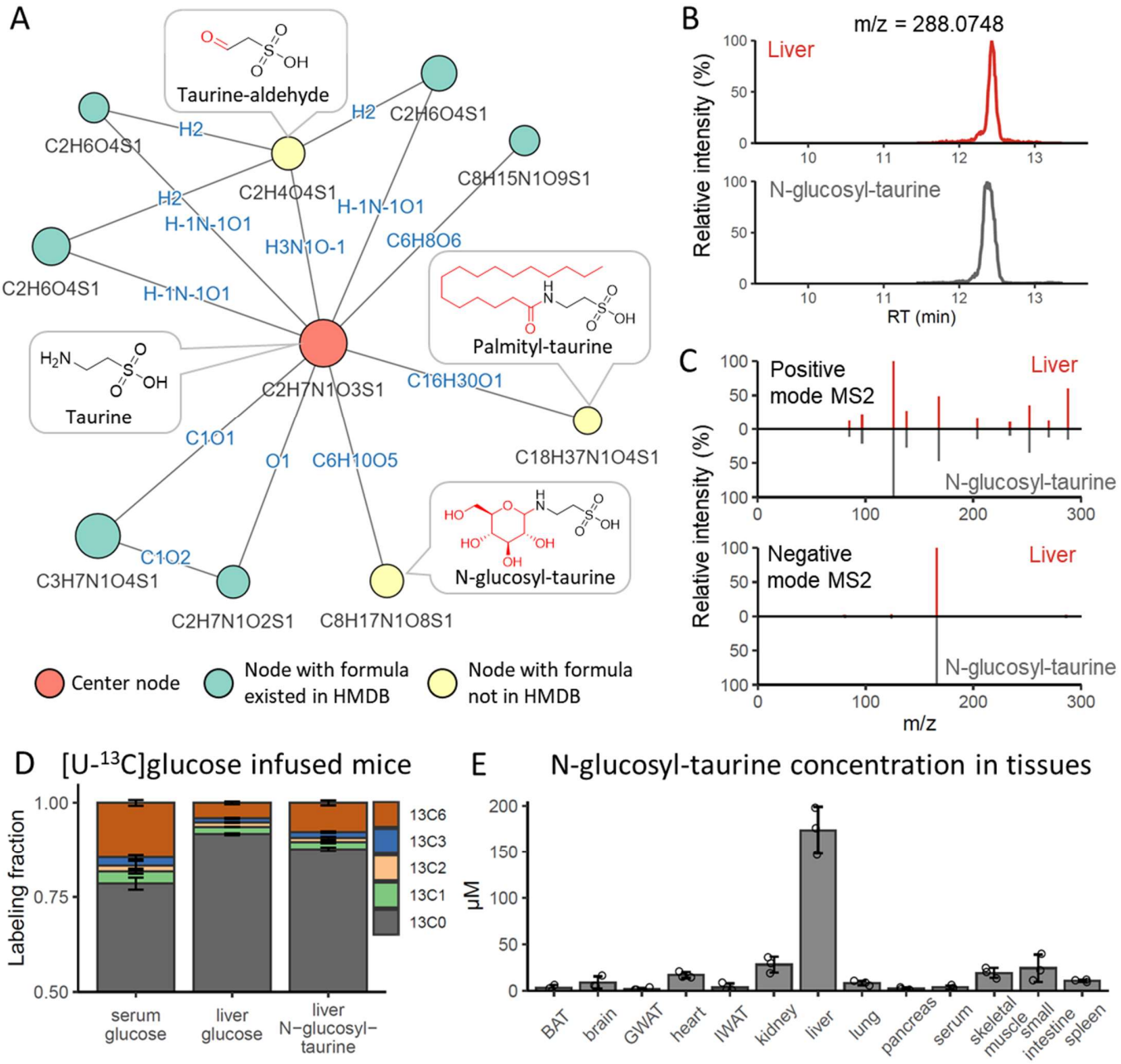


Figure 4. NetID discovers mammalian taurine derivatives. (A) Subnetwork surrounding taurine from mouse liver extract data. Nodes, connections, and formulae are direct output of NetID. Boxes with structures were manually added. (B) LC-MS chromatogram of N-glucosyl-taurine standard and the putative glucosyl-taurine from liver extract. (C) MS2 spectrum of glucosyl-taurine peak from liver extract (top), and synthetic N-glucosyl-taurine standard (bottom). (D) Isotope labeling pattern of putative glucosyl-taurine in mice, infused via jugular vein catheter for 2 h with [U-^{13}C]glucose. (E) Absolute N-glucosyl-taurine concentration in murine serum and tissues.

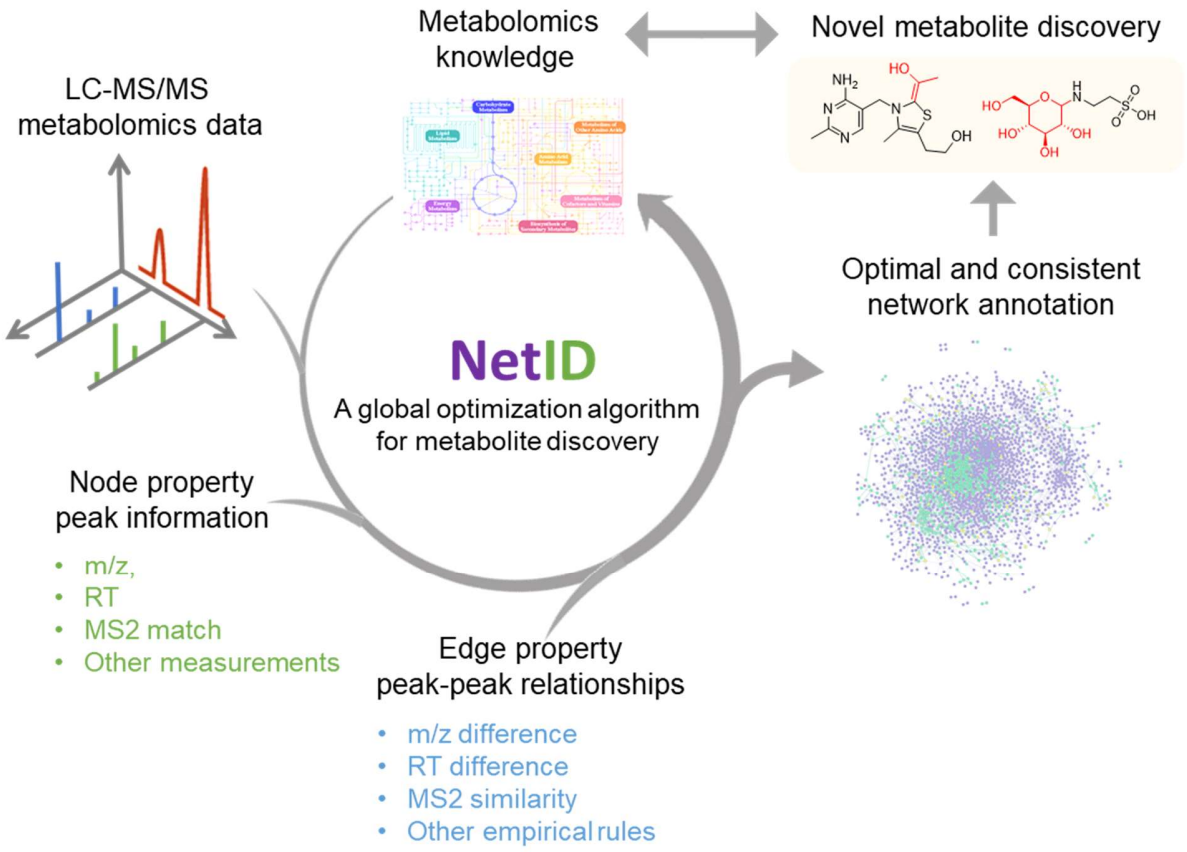


Figure 5. NetID applies global optimization for metabolomics data annotation and metabolite discovery.