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], Ziyang Chen[1,4], 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]Shanghai Key Laboratory of Metabolic Remodeling and Health, Institute of Metabolism & Integrative Biology, Fudan University, Shanghai, 200433, 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

Liquid chromatography-high resolution mass spectrometry (LC-MS)-based metabolomics aims to identify and quantitate all metabolites, but most LC-MS peaks remain unidentified. Here, we present a global network optimization approach, NetID, to annotate untargeted LC-MS metabolomics data. The approach aims to generate, for all experimentally observed ion peaks, annotations that match the measured masses, retention times, and (when available) MS/MS fragmentation patterns. Peaks are connected based on mass differences reflecting adducting, fragmentation, isotopes, or feasible biochemical transformations. Global optimization generates a single network linking most observed ion peaks, enhances peak assignment accuracy, and produces chemically-informative peak-peak relationships, including for peaks lacking MS/MS spectra. Applying this approach to yeast and mouse data, we identified five novel metabolites (thiamine derivatives and N-glucosyl-taurine). 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 the concentrations of detectable small-molecule in a biological system. In so doing, it reflects the integrated impact of genetics and 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 the 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. A common approach to peak annotation is to compare exact mass and either retention time or MS/MS (MS2) fragmentation pattern to authenticated standards. To facilitate such comparisons, extensive molecular structural databases (Pubchem[6], HMDB[7], KEGG[8], ChemSpider[9]), MS2 spectral databases (METLIN[10], GNPS[11], MassBank[12–14] and NIST[15]) and software (e.g. XCMS[16,17], GNPS[11], SIRIUS[18] and MS-DIAL[19]) have been developed. Peaks can also arise from mass spectrometry phenomena, such as adducts, fragments or isotopes of metabolites[20–24]. Such peaks seem to account for at least half of non-background LC-MS features[25–27]. Despite this progress, a great number of unknown peaks remain, and figuring out their identities is a primary challenge in the field.

Network analysis, capitalizing on peak-peak relationships to increase annotation scope and accuracy, has been broadly used in metabolomics data annotation. Workflows employing the concept of molecular connectivity have been used to build networks (e.g., GNPS[28–30], MetDNA[31], CliqueMS[32] and others[33–35]). Ions connected by either biochemistry or mass spectrometry phenomena often share MS2 fragmentation pattern similarity. While distinct metabolites typically separate chromatographically, ions connected through mass spectrometry phenomena co-elute.

Discovery of novel metabolites involves generation of candidate molecular formulae beyond those in current databases. This can be achieved by modifying formulae of known metabolites using characteristic biochemically feasible atom transformations that match observed MS1 mass differences[35,36] (e.g. 2.016 for 2H). Such approaches can be combined with clustering metabolites based on similar MS2 fragmentation patterns in a molecular network, as demonstrated in GNPS and other works[28–31,37]. In a cluster of connected peaks, one known metabolite peak can help to annotate its neighbors, facilitating unknown discovery. One state-of-the-art database-independent method to generate novel candidate molecular formulae (SIRIUS 4.0) combines high-resolution m/z, natural isotope abundances, and MS2 spectral analyses[18].

Existing methods generally focus on annotation of either individual peaks or a subnetwork of peaks. All peaks can be sequentially assessed, but individual annotations do not make full use of all available information regarding other peaks in the network. Global network optimization methods, not dealing with peak annotation one-by-one, but instead all at once to take full advantage of the entire available information, hold potential. For example, a Gibbs-sampling statistics approach has been used to incorporate biochemical connections, isotope patterns, and adduct relationships for probabilistic annotations of all peaks in a metabolomics experiment[38–40]. ZODIAC combines a similar approach and the SIRIUS algorithm[41]. Such Gibbs-sampling approaches allow network connections to shift the probabilities of candidate peak annotations, leading to substantially better annotations results[41]. While capable of assessing the probability of all individual candidate annotations in light of their neighbors, broadly defined, such approaches do not guarantee selection of the optimal overall network as a whole.

An alternative computational strategy focusing on optimizing the entire network, as opposed to statistical annotation of individual peaks, is integer linear programming. Given a scoring function, this computational optimization strategy determines the single best annotation for all peaks collectively. Global optimization of large networks of many thousands of peaks is feasible due to linear formulation of the problem, which renders the approach computationally efficient.

Integer linear programming optimization has, to our knowledge, not been previously applied in the context of molecular network analysis. To this end, we present an efficient stand-alone algorithm "NetID". The algorithm optimizes a network of mass spectrometry peak connections based on MS1 mass differences corresponding to gain or loss of relevant chemical moieties and MS2 spectral similarity, in a manner that differentiates biochemical connections from those based on mass spectrometry phenomena, and that incorporates literature data on known metabolites and their retention times. We applied this integer linear programming optimization approach to untargeted metabolomics data from both Baker's yeast and mouse liver. The global optimization step enforces a single formula assignment for each experimentally observed ion peak, increasing annotation accuracy as estimated by a target–decoy strategy[42]. Through these efforts, we provide likely formulae for several hundred putative novel metabolites and confirm the identities of five new metabolites.

Results

NetID algorithm

NetID involves three computational steps: candidate annotation, scoring, and network 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 candidate annotation phase, we match every experimentally measured node m/z to formulae in the selected metabolomics database (e.g.

HMDB). Peaks matching to database formula within 10 ppm are assigned as seed nodes with candidate seed formulae, 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, we make candidate edge annotations such that (i) $\Delta m/z$ 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 starting from the seed nodes, candidate formulae are assigned to nodes outside the initial seeds. A few rounds of edge extension suffice to give thorough coverage (see Methods). Due to finite mass measurement precision, a single node (including a seed node) may be assigned multiple contradictory candidate formulae, which are resolved at the following scoring and optimization step.

NetID then scores every candidate node and edge annotation. Candidate node annotations are scored based on precision of m/z match to the molecular formula and (when the relevant information is available) precision of retention time match to known metabolite retention time and quality of MS2 spectra match to database structure. In addition, there is a bonus for matching to formula in HMDB and a penalty for unlikely formulae (e.g. containing an uncommon elemental ratio or extreme number of ring and double bond equivalents)[43]. Biochemical edges receive a positive score for MS2 spectra similarity between the connected nodes. 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 and Supplementary Note 2). The overall impact is to assign high scores to those candidate annotations that effectively align the experimentally observed ion peaks with prior metabolomics knowledge.

With a score assigned for each candidate 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 annotation and that they 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 minutes to hours 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, assigning 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 from seed assignments. 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 (mis)annotated as a putative novel 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.

We applied the NetID algorithm to yeast and liver datasets, in both positive and negative ionization mode (Figure 2B, Supplementary Fig. 1A). Raw LC-MS data from replicate yeast or liver samples were analyzed together by peak-picking software (El-MAVEN[44]) to generate a single list of peaks consistently found for that sample type and ionization mode. Yeast data were MS1 only, while liver data included targeted MS2 spectra. Considering the example of negative mode yeast data with a total of 5,588 non-background peaks, in the candidate 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 (formula corresponding to $M \pm H$ monoisotopic peak existed in database), 14% were putative novel metabolites (formula not in database but with biochemical connection to a metabolite), and the rest were mass spectrometry phenomena, such as adducts, fragments, isotopes. Thus, after thorough background ion removal, we assign a few thousand peaks as likely metabolite ions and the majority as mass spectrometry artifacts. Orthogonal approaches such as credentialing via isotope labeling[27] similarly assign the majority of peaks as mass spectrometry artifacts, but annotate fewer peaks as likely metabolites than NetID.

The roughly 5000 nodes were connected by about 10,000 edges. Two nodes share each edge, with each node connected by an average of four edges. These edges were roughly evenly split between biochemical and abiotic connections (Figure 2B, Supplementary Fig. 1A,B). More than 90% of annotated nodes fell into a single dominant connected network (Supplementary Fig. 1C), reflecting most peaks being connected to core metabolism. About 15% of peaks (737/5588), however, remained unannotated (Figure 2B). 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 metabolites (Supplementary Table. 3). The annotated peaks included several hundred formulae for putative novel metabolites (Supplementary Fig. 2, Supplementary Data 1).

Performance validation

We evaluated the performance of the NetID algorithm using the negative mode yeast data (MS1 only). We first employed a target-decoy estimation strategy, in which we intentionally introduce formulae with biologically unreasonable elements, and test whether our annotation strategy effectively avoids annotating peaks with these fake formulae[45,46]. Assessments were made using a number of different metabolite databases (HMDB[7], YMDB[47], PubChemLite.0.2.0[48], and a subset of biopathway related entries in PubChemLite.0.2.0). As expected, the smaller databases yielded fewer false identifications. Importantly, across all of the databases, NetID more effectively selected appropriate formulae (lower false discovery rate) compared to methods considering m/z only, node scores only, or both node and edge scores but without global optimization (Figure 2C, Supplementary Fig. 3A).

As an orthogonal means of testing the algorithm, we manually curated 314 peaks as known annotations (Supplementary Data 1), and assessed the fraction annotated correctly. Across databases, NetID resulted in more accurate annotations of these gold standard peaks, with the number of incorrect annotations roughly an order of magnitude lower for NetID compared to node or combined node and edge scores without global optimization (Figure 2C, Supplementary Fig. 3B).

Thiamine-derived metabolites

Among the putative novel metabolites in the yeast metabolomics dataset, we found three with total ion current > 10^5 that are 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, Supplementary Fig 4). These formula assignments and connections were initially obtained without MS2 spectra being available, reflecting the ability of NetID to make accurate formula assignments and connections based on MS1 data (combined with other peak attributes like retention time). 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.

MS2 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 confirmed 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, while the addition of C_2H_4O is much less common biochemically. 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 of carbohydrate to TCA cycle) (Figure 3E). The de-pyrophosphorylation product of thiamine pyrophosphate intermediate in PDH reaction yields thiamine+C_2H_4O (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. 5). Thus, NetID enabled the discovery of four novel thiamine-derived metabolites that were not present in metabolomics databases (Supplementary Table 4).

N-glucosyl-taurine

We similarly carried out NetID annotation of a mouse liver dataset. We observed multiple putative novel 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, Supplementary Fig. 6). Like the thiamine-related peaks, these were initially correctly annotated without relying on MS2 data. While all were missing in HMDB, the latter two were found in METLIN: N-palmitoyl taurine ($C_{18}H_{37}NO_4S$) and sulfoacetaldehyde ($C_2H_4O_4S$). Pubchem contains an entry for N-glucosyl-taurine ($C_8H_{17}NO_8S$) as a synthetic chemical but no database previously suggested it being a metabolite. To confirm 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 MS2 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 biosynthesis of N-glucosyl-taurine from circulating glucose (Figure 4D). N-glucosyl-taurine was not observed in yeast extract but was detected in multiple mouse tissues. Search for peaks matching the N-glucosyl-taurine MS2 spectra using MASST identified matches in both mouse and human samples[49]. Quantitation using the synthetic standard shows that liver has the highest level of glucosyl-taurine at ~170 µM (Figure 4E, Supplementary Fig. 7). 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 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 three mass peaks for every metabolite or putative metabolite (e.g. M+H plus two 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. The number of actual novel metabolites being detected could potentially be more, due to isomers sharing the same formula as known metabolites, or 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 five species for which we assign structures (Figure 3, 4).

A key benefit of molecular network-based annotation is the ability to assimilate steadily new information[28,29]. MS2 data, which can be collected in either targeted or data-dependent mode, provide fragmentation information to facilitate peak annotation. Due to the global network optimization approach, new MS2 spectra facilitate not only annotation of the ions being fragmented, but also other peaks throughout the network. Other data types can be seamlessly added. For example, compound class categorization based on MS2 data[50] or retention time prediction[51,52] can be added to score nodes. Isotope labeling similarity upon feeding different "heavy" nutrients could potentially be added to score edges. Global optimization integrates new information comprehensively with prior knowledge to arrive at optimal annotations. Cycles of experimentation and machine learning hold the potential to identify most unknown metabolites over the coming decade, eventually 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[53]. 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 least 15 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 standard mouse chow 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[54]. 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.

Glucosyl-taurine synthesis

Glucosyl-taurine synthesis was carried out following previous literature reports with slight modifications[55]. 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 min 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 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[56] 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 K 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 mixed with unreacted substrates (Supplementary Fig. 8).

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.

MS2 spectra were collected in targeted mode using the PRM function at 25 eV HCD energy with other instrument settings being resolution 17500, AGC target 10^6, Maximum IT 250 ms, isolation window 1.5 m/z. Targeted MS2 data for the thiamine related metabolites were collected for structural identification using similar parameters as above except the HCD energy were set at 20, 35, and 50 eV in a step-CE mode.

Data prepossessing

LC-MS raw data files (.raw) were converted to mzXML format using ProteoWizard[57]. El-MAVEN (version 7.0) was used to generate a peak table containing m/z, retention time, and 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.

Targeted MS2 data were extracted from the mzXML files using lab-developed Matlab code (Supplementary Note 1). 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 data input. Although the provided workflow uses targeted MS2 data as input, NetID as currently configured can also handle data-dependent MS2 data, but additional parsing software (under development) is need to convert the large primary data-dependent MS2 files into the NetID input format.

Data input

NetID algorithm requires three types of input files: a peak table (in .csv format) recording m/z, retention time, and intensity for peaks; an atom difference rule table (in .csv format), which in our case consisted of 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 structures, formulae, m/z and MS2 spectra of metabolites from the selected metabolomics database (e.g. HMDB) including retention times of known metabolites under the relevant LC conditions. MS2 data input is optional. For the analysis of yeast and mouse liver datasets in Figure 2-4, 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 (Supplementary data 2). For yeast, no MS2 were used in NetID analysis (MS2 were used post hoc to confirm certain annotations). For liver, targeted MS2 spectra were used (1479 positive and 803 negative ionization mode spectra experimentally obtained for previously identified peaks of > 10^5 intensity[58]).

Candidate node and edge annotations

The first module of NetID algorithm is to make candidate annotations for seed nodes, assign candidate annotations for other nodes, and assign candidate 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 selected metabolomics database (e.g. HMDB). When the m/z difference is within a predefined tolerance (e.g. 10 ppm), candidate formulae and IDs are assigned to the node, and this node is defined as a primary seed node. Note that assignments to seeds are candidate annotations. A primary seed node can contain multiple candidate formulae and IDs if all are within the m/z difference range.

Edges connect two nodes via gain or loss of specific atoms. We provided a list of 25 biochemical atom differences and 59 abiotic atom differences which together define all connections in the network (Supplementary Table 1, 2, Supplementary Data 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 within a predefined m/z_{tol} (e.g. 10 ppm), 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

$$| \left(v_{\mathrm{m/z}} - u_{\mathrm{m/z}}\right) - D_i | < v_{\mathrm{m/z}} \times m/z_{tol} \quad (3)$$

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 a predefined RT_{tol} (e.g. 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

$$| v_{\mathrm{RT}} - u_{\mathrm{RT}} | < RT_{tol} \tag{4}$$

For each node, its candidate formulae set will expand due to extending formulae from its neighboring nodes through the 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 a predefined m/z tolerance (e.g. 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 themselves can 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.

Each candidate node annotation is defined as (i) metabolite, (ii) putative novel metabolite, or (iii) artifact (nodes can also be unannotated). Specifically, if the elemental formula corresponding to the (de)protonated ion of a monoisotopic peak is found in the employed metabolomics database, this node is defined as a metabolite. If the formula is not found at the employed database, but the node is connected to a metabolite via biochemical connection(s), the node is defined as a putative novel metabolite. Finally, if the node connected only via abiotic connections such as adduct, fragment, or isotope connection(s), it is defined as an artifact. As currently configured, NetID defines metabolite peaks exclusively as (de)protonated ions. In the case that a (de)protonated ion peak is not detected, but related adducts are (e.g. $[M+Na]^+$), the adducts will remain unannotated (or be misannotated), as there is no procedure for annotating adducts lacking (de)protonated ion peaks.

Scoring in NetID

The second module of NetID algorithm is to score every candidate node and edge annotation assigned in the candidate 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, MS2, 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 $\mathrm{S}(u, a_i)$ denotes the score of candidate annotation a_i for node *u*. $\mathrm{S}(u, a_i)$ is the sum of seven different scoring components, including (a) $\mathrm{S}_{\mathrm{m/z}}$, a negative score evaluating the difference of between measured m/z and the calculated m/z of assigned molecular formula; (b) S_{RT}, a positive score if the measured RT for the peak corresponding to node *u* matches to a known standard; (c) $\mathrm{S}_{\mathrm{MS2}}$, a positive score if the measured MS2 spectrum of node *u* matches the database MS2 spectrum of annotation a_i; (d) $\mathrm{S}_{\mathrm{database}}$, a positive score if the annotated formula a_i exists in the employed metabolomics database; (e) $\mathrm{S}_{\mathrm{missing_isotope}}$, a negative score if an expected isotopic peak is missing; (f) $\mathrm{S}_{\mathrm{rule}}$, a negative score if annotation a_i violates basic chemical rules; (g) $\mathrm{S}_{\mathrm{derivative}}$, a positive score if the annotation a_i is derived from a parent peak with a high score annotation. For details, see Supplementary Note 2.

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. Biochemical, isotope, adduct, and neural loss edge annotations are the most common types. We also score other less common abiotic connection types appeared in orbitrap data, including oligomers, multi-charge species, heterodimers, in-source fragments of known or unknown metabolites[59], and ringing artifact peaks surrounding high intensity ions[26,60].

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.

$$\mathrm{S}(u, v, a_i, b_j, D_k) = 0, \text{ if } |a_i - b_j| \neq D_k \tag{5}$$

$\mathrm{S}(u, v, a_i, b_j, D_k)$ is the sum of four different scoring components, including (a) $\mathrm{S}_{\mathrm{MS2_similarity}}$, a positive score defined for biochemical edges if node *u* and *v* have experimental measured MS2 spectra, and they share MS2 similarity; (b) $\mathrm{S}_{\mathrm{co_elution}}$, a negative score defined for abiotic edges, if the RT of two connected nodes differ more than a threshold (e.g. 0.05 min); (c) $\mathrm{S}_{\mathrm{type}}(u, v, a_i, b_j, D_k)$, a non-negative score defined for all edges, depending on the connection type of edge, which is defined by D_k, including biotransformation, adduct, isotope, and fragment edges, and optionally

including oligomer and multi-charge species, heterodimer, in-source fragments and ringing artifacts edges; (d) $S_{isotope_intensity}$, a negative score defined for isotopic edges (a type of abiotic edge) if the measured isotope peaks deviate from expected natural abundance. For details, see Supplementary Note 2.

Global network optimization using linear programing

The third module of NetID algorithm is to perform global network optimization. Using scores assigned for each candidate node and edge annotation, our goal is to find formula annotations for each node and edge so as to maximize the sum of their scores across the network under the constraints that each node is assigned a single annotation, and that the network annotation is consistent. When a node has multiple candidate node annotations shared with same formula (e.g. isomers), the one with highest score (better MS2 match or RT match) is selected. When equal scores happen, the candidate annotation that appears first in the metabolite list from database is reported as a default. We use linear programming to solve this optimization problem, 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,

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

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

For each edge, 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,

$$c_{u,v,a_i,b_j,D_k} = 1, \text{ if } |a_i - b_j| = D_k$$

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

We constrain the optimization so that each node has a single annotation, and an edge exists only if the atom difference of that edge annotation matches the formula difference of nodes. For computational purposes, nodes may also receive “blank” or “no formula” annotation (we refer to such nodes in elsewhere in the text as “unannotated”). The node and edge binary variables must 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 1 or 0, representing the candidate annotation is selected or not selected, respectively, in the network.

With scores of each candidate node and edge annotation, the objective for the optimization is to determine 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 above 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. For the reported datasets and using the default parameter settings, optimization finishes within an hour on a standard laptop (Supplementary table 5). Depending on the number of peaks in data tables, the entries in the atom difference tables, the choice of reference compound databases, and the parameters involved in scoring, runtimes during internal testing ranged from minutes to hours.

Evaluation of NetID

After running the candidate annotation and network extension process as usual for NetID, we compared four different annotation selection methods: (i) m/z only, selecting the candidate annotation with closest m/z to the measured m/z; (ii) node scores, selecting the candidate annotation with highest candidate node score (as per usual NetID scoring rules); (iii) node + edge scores, adding half of a candidate edge score to each node score of the two connected candidate node annotations, and selecting the candidate annotation with highest combined score; (iv) NetID optimization, using the candidate node and edge score, and selecting the candidate annotation from global optimization.

We employed a target-decoy strategy to estimate false discovery rate[45,46]. The target library is the compound library we use for annotation, including HMDB (human metabolomics database), PBCM (PubChemLite.0.2.0)[48], PBCM_BIO (a subset of biopathway related entries in PubChemLite.0.2.0) and YMDB (yeast metabolomics database)[47]. The decoy formula was generated by adding an implausible element adduct to a formula from target library. These implausible elements are those not in any formulae in database, namely, He, Be, Ne, Sc, Kr, Rb, Sr, Y, Zr, Nb, Mo, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, Xe, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Th and U. Applying the decoy formula generation process (i.e. adding a single

randomly selected implausible element in place of hydrogen) to all unique formulae in a target compound library gives the corresponding decoy library. The combined target library and decoy library were used for annotation. Any annotation containing an implausible element is considered a false positive (FP) result. The number of FP results from target library is estimated to be similar to that from decoy library, because 1:1 ratio of target formulae and decoy formulae were used. That is $FP_{target_library} \approx FP_{decoy_library}$. Using the combined target-decoy library, the false discovery rate (FDR) is estimated to be $FP_{decoy_library} / (T - FP_{decoy_library})$, where T is the total number of annotations. The decoy library generation process was repeated ten times for each database.

We manually annotated 314 peaks in the yeast negative mode dataset (Supplementary data 1). Using these annotations as a ground truth, we evaluated the fraction of correct annotation for the four different annotation methods above. Peak annotations matching to the ground truth formulae, including adduct, isotope formulae, are counted as correct, and peaks that are not annotated or their annotations did not match are counted as incorrect. The annotation process used 1:1 target-decoy library and was repeated 10 times as above.

Visualization

We provide an interactive Shiny R app to visualize and explore the NetID output network. In addition, NetID outputs two .csv files (cyto_node.csv and cyto_edges.csv) that are compatible with the general network visualization software Cytoscape. The interactive Shiny R app and detailed user guide are available at GitHub (https://github.com/LiChenPU/NetID).

Data availability

All raw data, including the yeast, the mouse studies, the ^{13}C studies and over 2000 targeted MS2 files collected from the liver data in mzXML formats were deposited in MassIVE (ID = MSV000087434). MS2 spectra of reported novel metabolites are included in Supplementary Data 3. R code for generating NetID statistics and performing FDR analysis in Figure 2 and Supplementary Fig 1 are provided in GitHub (https://github.com/LiChenPU/NetID). Peak table for yeast data negative mode, including NetID annotation results, putative novel metabolite list, and manual curation results are provided in Supplementary Data 1. Exemplary peak table from the yeast dataset, atom difference rule table, HMDB compound list (customized to contain relevant information), and an in-house retention time list for known metabolites are provided in Supplementary Data 2.

Code availability

NetID was developed mainly in R, and used a mixture of IBM ILOG CPLEX Optimization Studio, Matlab and Python. NetID code and example files are available for non-commercial use in GitHub at https://github.com/LiChenPU/NetID, under the GNU General Public License v3.0. User guide, parameter explanation and pseudocode are provided in Supplementary Notes 1, 2, and 3.

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 at Princeton University for the NMR analysis, the Metabolomics & Lipidomics Mass Spectrometry Core Facility of IMIB at Fudan University for additional mass spectrometry support, 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., X.X. and Z.C. wrote the NetID algorithm code. 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. analyzed 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. synthesized 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 with unique formulae and connected by edges reflecting atom differences arising either through metabolism (biochemical connection) or mass spectrometry phenomenon (abiotic connection). Peaks are classified as "metabolite" (M+H or M-H peak of formula found in selected metabolomics database, e.g. HMDB), "putative novel metabolite" (formula not found in database but with biochemical connection to a metabolite), or "artifact" (only abiotic connection to a metabolite). NetID algorithm involves three steps. Candidate annotation first matches peaks to database 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 database 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) Summary table of NetID annotations of negative and positive mode LC-MS data from Baker's yeast and mouse liver. (C) False discovery rate estimated using target-decoy strategy. Each individual data point (circle) is from a different randomized decoy library (n = 10 randomized libraries were tested). (D) Fraction of 314 manually curated "ground truth" annotations made correctly (using the 10 libraries). Boxes show median and interquartile range.

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 (n = 5). (D) The thiamine derivatives are also found in mouse tissues and urine (n=3). (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. Bar represents mean values and error bar indicates s.d. in (C) and s.e. in (D).

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) Top 10 abundant ion peaks in 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 (n=3). (E) Absolute N-glucosyl-taurine concentration in murine serum and tissues (n=3). Bar represents mean values and error bar indicates s.d. in (D) and s.e. in (E).

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. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* **47**, D1102–D1109 (2019).
7. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46**, D608–D617 (2018).
8. 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).
9. Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **87**, 1123–1124 (2010).
10. Xue, J., Guijas, C., Benton, H. P., Warth, B. & Siuzdak, G. METLIN MS2 molecular standards database: a broad chemical and biological resource. *Nat Methods* **17**, 953–954 (2020).

11. 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).
12. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.* **88**, 7946–7958 (2016).
13. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
14. MassBank | MassBank Europe Mass Spectral DataBase. https://massbank.eu/MassBank/.
15. sherena.johnson@nist.gov. NIST Standard Reference Database 1A. *NIST* https://www.nist.gov/srd/nist-standard-reference-database-1a (2014).
16. 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).
17. Forsberg, E. M. *et al.* Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature Protocols* **13**, 633–651 (2018).
18. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16**, 299–302 (2019).
19. Tsugawa, H. *et al.* A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods* **16**, 295 (2019).
20. Stricker, T., Bonner, R., Lisacek, F. & Hopfgartner, G. Adduct annotation in liquid chromatography/high-resolution mass spectrometry to enhance compound identification. *Anal Bioanal Chem* **413**, 503–517 (2021).
21. 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).
22. Domingo-Almenara, X. *et al.* Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. *Anal. Chem.* **91**, 3246–3253 (2019).
23. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).
24. Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P. & Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **90**, 480–489 (2018).
25. Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.* **142**, 9097–9105 (2020).
26. Wang, L. *et al.* Peak Annotation and Verification Engine for Untargeted LC–MS Metabolomics. *Anal. Chem.* **91**, 1838–1846 (2019).

27. Mahieu, N. G., Huang, X., Chen, Y.-J. & Patti, G. J. Credentialing Features: A Platform to Benchmark and Optimize Untargeted Metabolomic Methods. *Anal. Chem.* **86**, 9583–9589 (2014).
28. 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.
29. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* **17**, 905–908 (2020).
30. da Silva, R. R. *et al.* Propagating annotations of molecular networks using in silico fragmentation. *PLOS Computational Biology* **14**, e1006089 (2018).
31. Shen, X. *et al.* Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nature Communications* **10**, 1516 (2019).
32. 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.
33. Alden, N. *et al.* Biologically Consistent Annotation of Metabolomics Data. *Anal. Chem.* **89**, 13097–13104 (2017).
34. 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.
35. Yu, M. & Petrick, L. Untargeted high-resolution paired mass distance data mining for retrieving general chemical relationships. *Commun Chem* **3**, 157 (2020).
36. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *PNAS* **109**, E1743–E1752 (2012).
37. Hooft, J. J. J. van der, Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *PNAS* **113**, 13738–13743 (2016).
38. 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.* **91**, 12799–12807 (2019).
39. Rogers, S., Scheltema, R. A., Girolami, M. & Breitling, R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* **25**, 512–518 (2009).
40. Daly, R. *et al.* MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **30**, 2764–2771 (2014).
41. Ludwig, M. *et al.* Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence* **2**, 629–641 (2020).
42. Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods* **14**, 57–60 (2017).

43. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **8**, 105 (2007).
44. Melamud, E., Vastag, L. & Rabinowitz, J. D. Metabolomic Analysis and Visualization Engine for LC−MS Data. *Anal. Chem.* **82**, 9818–9826 (2010).
45. Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Methods* **14**, 57–60 (2017).
46. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* **7**, 29–34 (2008).
47. Jewison, T. *et al.* YMDB: the Yeast Metabolome Database. *Nucleic Acids Research* **40**, D815–D820 (2012).
48. Bolton, E. & Schymanski, E. PubChemLite tier0 and tier1. (2020) doi:10.5281/zenodo.3611238.
49. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat Biotechnol* 1–4 (2020) doi:10.1038/s41587-019-0375-9.
50. 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.
51. 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).
52. Bach, E., Szedmak, S., Brouard, C., Böcker, S. & Rousu, J. Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics* **34**, i875–i883 (2018).
53. Xu, Y.-F. *et al.* Discovery and Functional Characterization of a Yeast Sugar Alcohol Phosphatase. *ACS Chem. Biol.* **13**, 3011–3020 (2018).
54. Hui, S. *et al.* Glucose feeds the TCA cycle via circulating lactate. *Nature* **551**, 115–118 (2017).
55. 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).
56. 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).
57. Chambers, M. C. *et al.* A Cross-platform Toolkit for Mass Spectrometry and Proteomics. *Nat Biotechnol* **30**, 918–920 (2012).
58. 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).
59. 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).

60. 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).

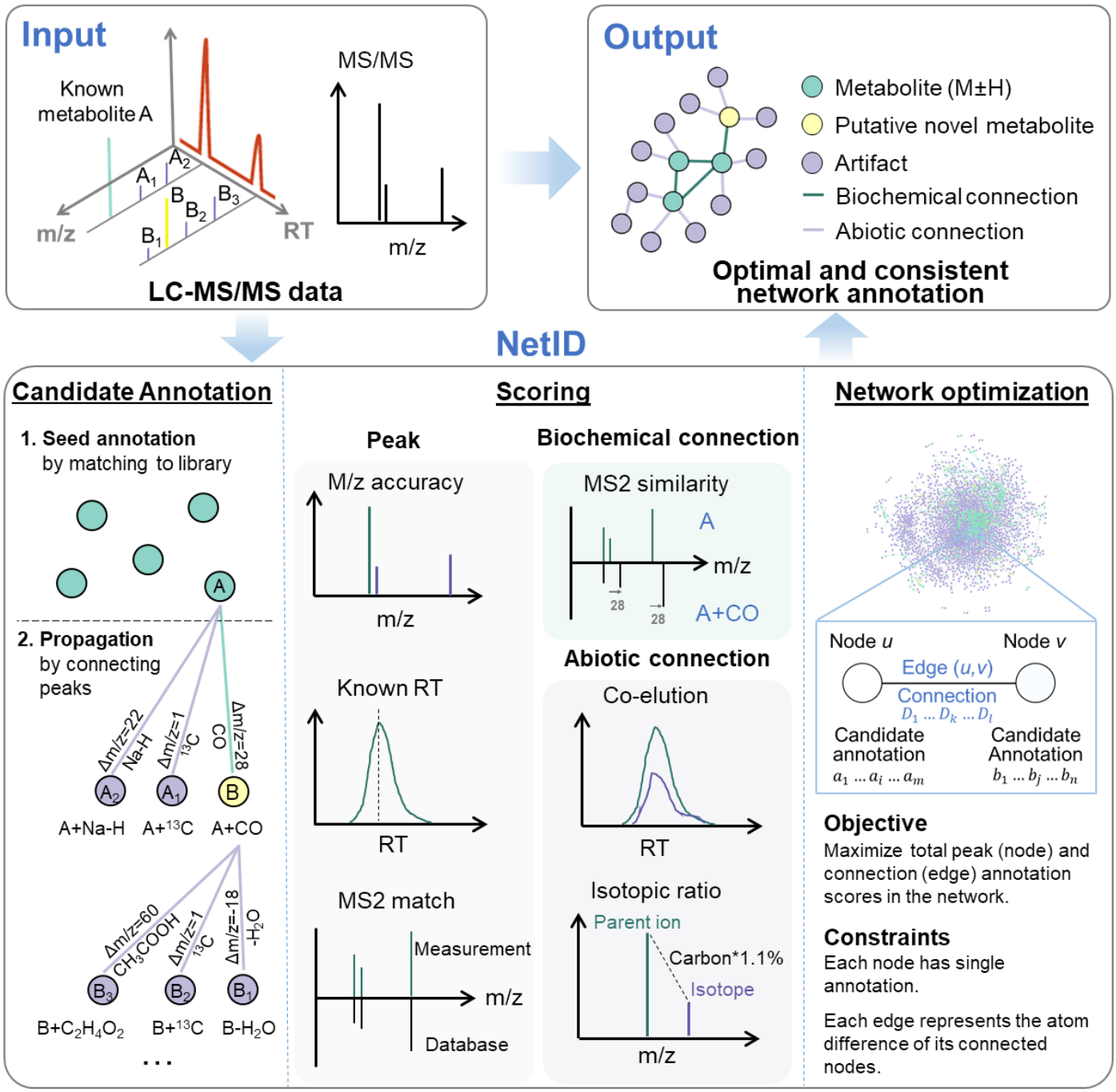
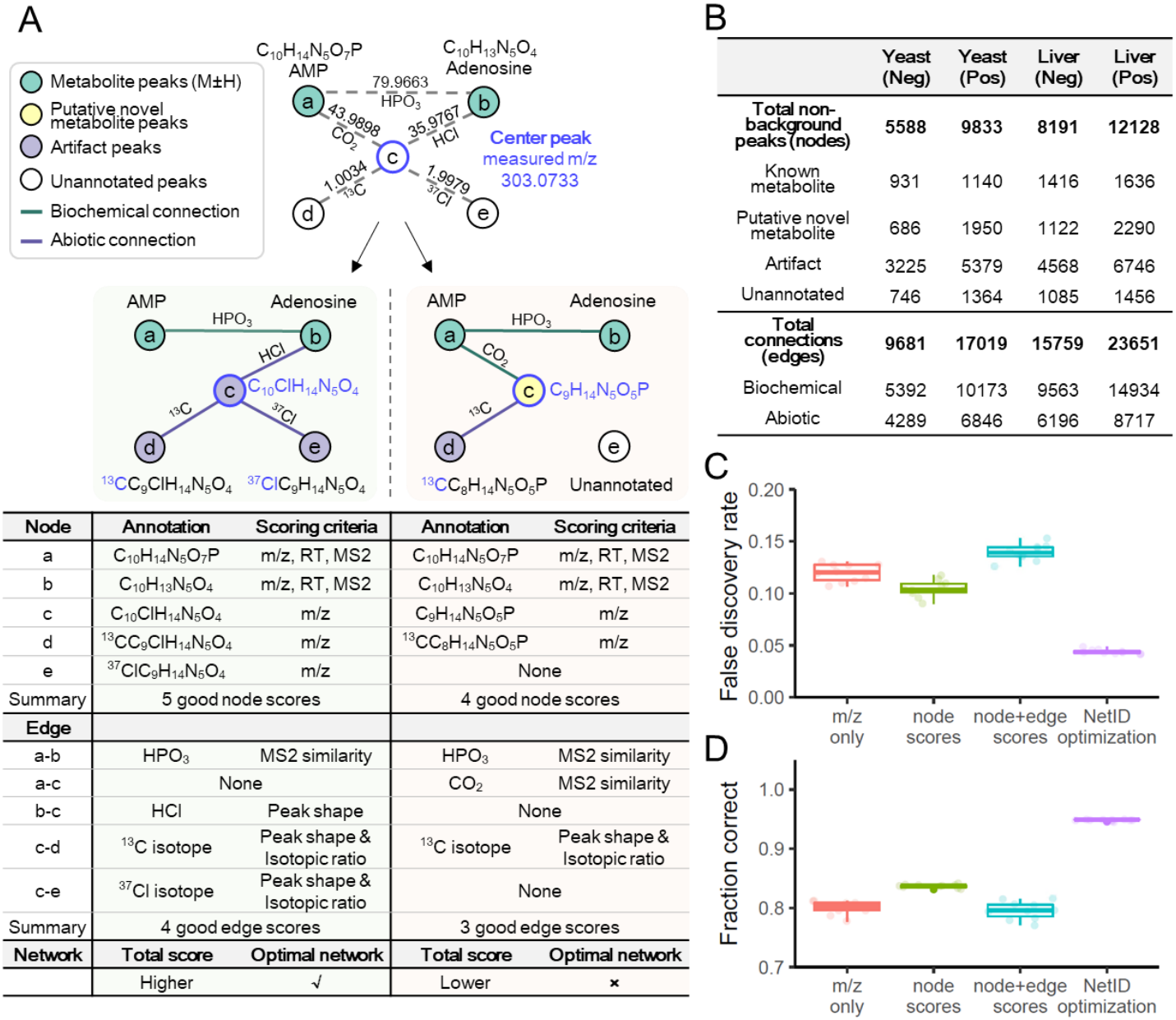


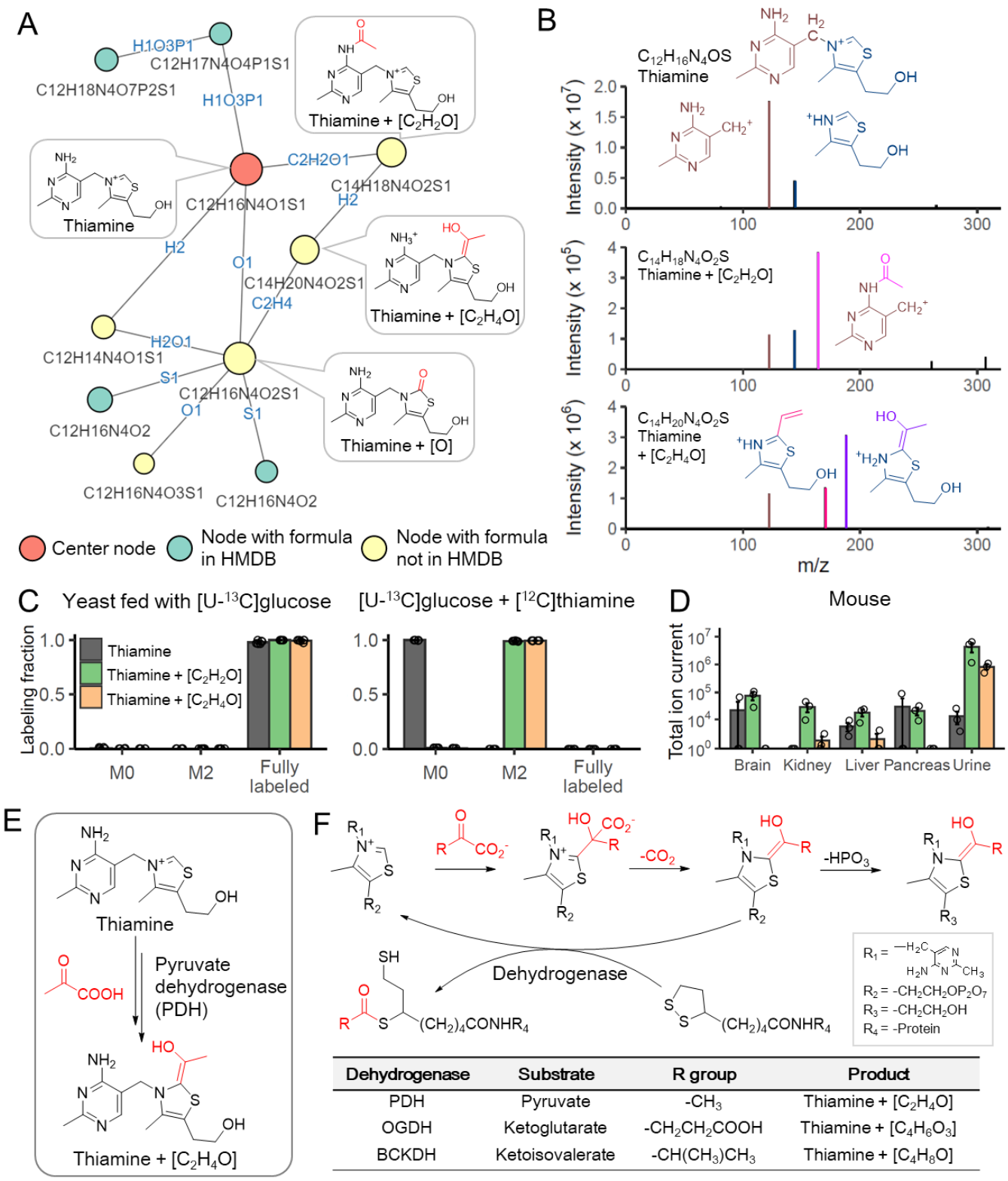
Figure 1. A global network optimization approach for untargeted metabolomics data annotation (NetID). The input data are LC-MS peaks with m/z, retention times, intensities and optional MS2 spectra. The output is a molecular network with peaks (nodes) assigned with unique formulae and connected by edges reflecting atom differences arising either through metabolism (biochemical connection) or mass spectrometry phenomenon (abiotic connection). Peaks are classified as "metabolite" (M+H or M-H peak of formula found in selected metabolomics database, e.g. HMDB), "putative novel metabolite" (formula not found in database but with biochemical connection to a metabolite), or "artifact" (only abiotic connection to a metabolite). NetID algorithm involves three steps. Candidate annotation first matches peaks to database 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.



	Yeast (Neg)	Yeast (Pos)	Liver (Neg)	Liver (Pos)
Total non-background peaks (nodes)	**5588**	**9833**	**8191**	**12128**
Known metabolite	931	1140	1416	1636
Putative novel metabolite	686	1950	1122	2290
Artifact	3225	5379	4568	6746
Unannotated	746	1364	1085	1456
Total connections (edges)	**9681**	**17019**	**15759**	**23651**
Biochemical	5392	10173	9563	14934
Abiotic	4289	6846	6196	8717

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	×

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 database 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) Summary table of NetID annotations of negative and positive mode LC-MS data from Baker's yeast and mouse liver. (C) False discovery rate estimated using target-decoy strategy. Each individual data point (circle) is from a different randomized decoy library (n = 10 randomized libraries were tested). (D) Fraction of 314 manually curated "ground truth" annotations made correctly (using the 10 libraries). Boxes show median and interquartile range.



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 (n = 5). (D) The thiamine derivatives are also found in mouse tissues and urine (n=3). (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. Bar represents mean values and error bar indicates s.d. in (C) and s.e. in (D).

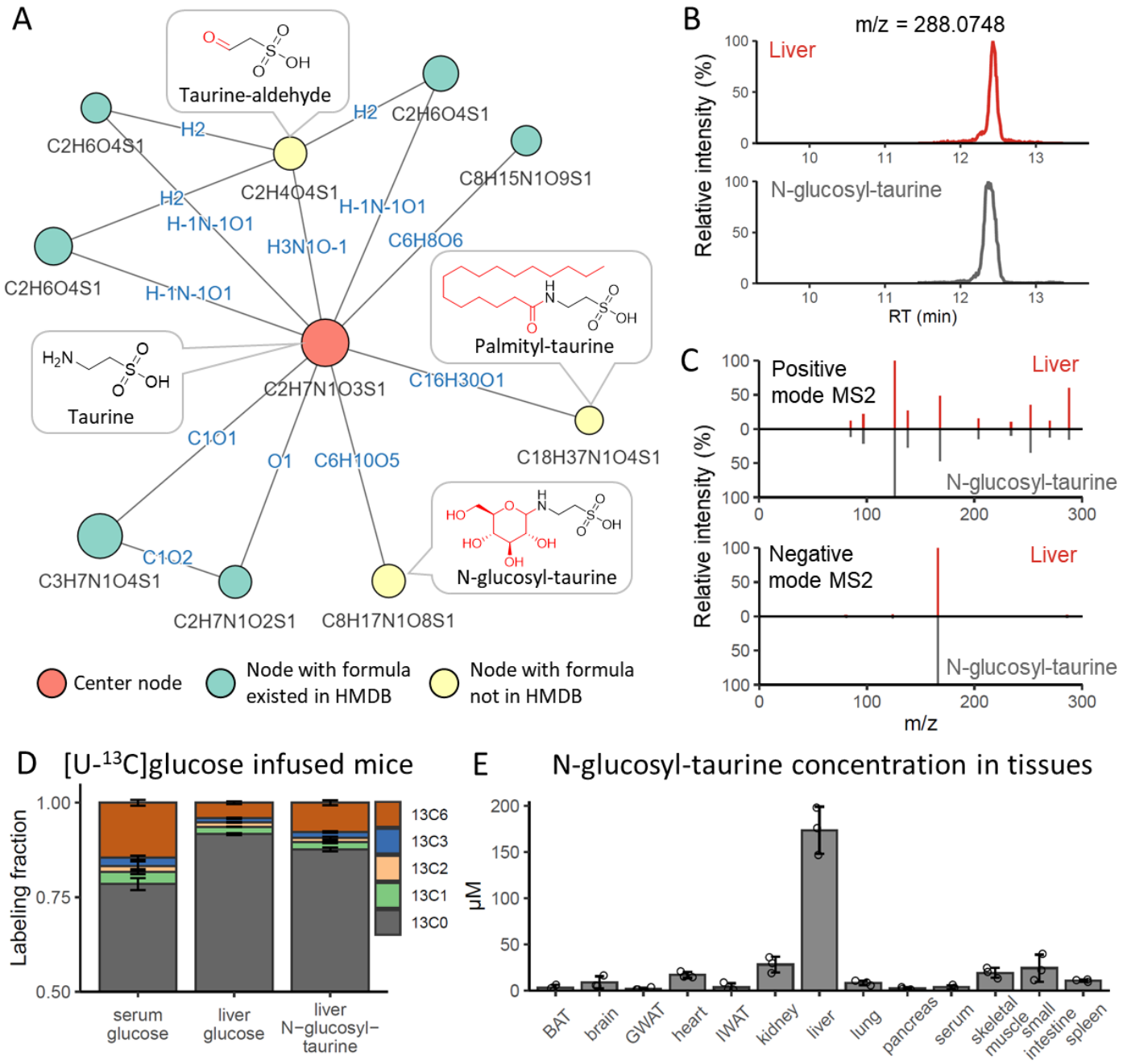


Figure 4. NetID discovers mammalian taurine derivatives. (A) Subnetwork surrounding taurine from mouse liver extract data. Nodes, connections, and formulae are direct output of NetID. Boxes with structures were manually added. (B) LC-MS chromatogram of N-glucosyl-taurine standard and the putative glucosyl-taurine from liver extract. (C) Top 10 abundant ion peaks in 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 (n=3). (E) Absolute N-glucosyl-taurine concentration in murine serum and tissues (n=3). Bar represents mean values and error bar indicates s.d. in (D) and s.e. in (E).

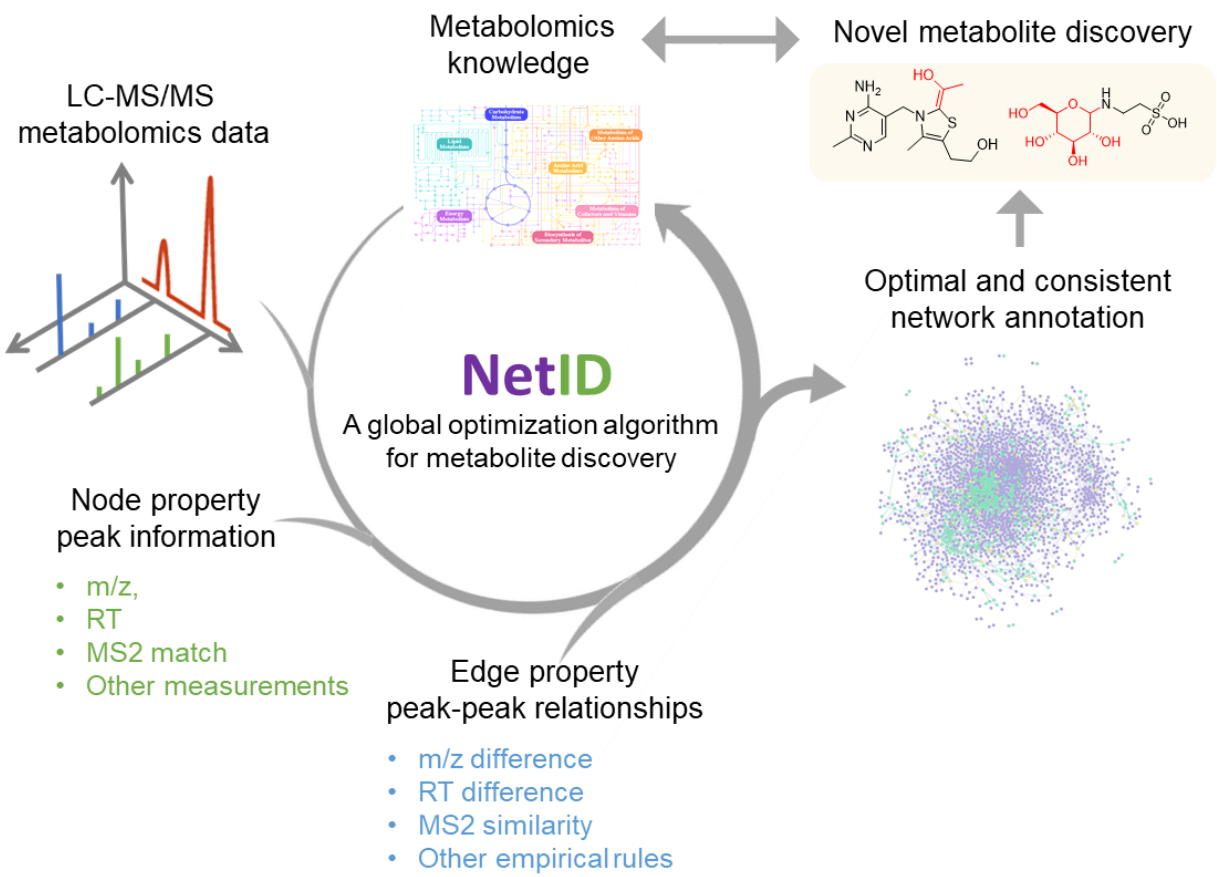


Figure 5. NetID applies global optimization for metabolomics data annotation and metabolite discovery.