### Title

Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets




### Authors

Barupal, Dinesh Kumar
Fan, Sili
Fiehn, Oliver



### Publication Date

2018-12-01

### DOI

10.1016/j.copbio.2018.01.010

Peer reviewed

HHS Public Access

Author manuscript

*Curr Opin Biotechnol*. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

*Curr Opin Biotechnol*. 2018 December ; 54: 1–9. doi:10.1016/j.copbio.2018.01.010.

# Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets

**Dinesh Kumar Barupal**[1], **Sili Fan**[1], and **Oliver Fiehn**[1,2]

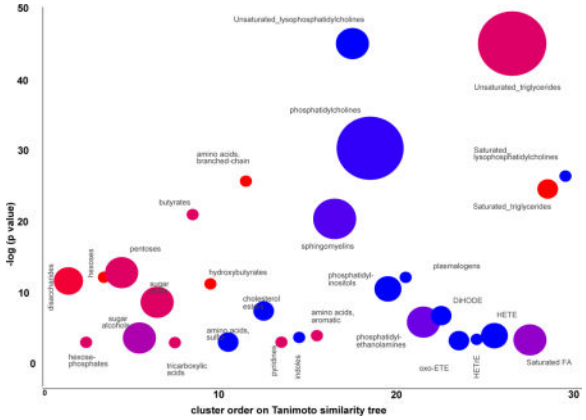
[1]NIH West Coast Metabolomics Center, University of California Davis, Davis 95616, CA, United States

[2]Biochemistry Department, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia

## Abstract

Access to high quality metabolomics data has become a routine component for biological studies. However, interpreting those datasets in biological contexts remains a challenge, especially because many identified metabolites are not found in biochemical pathway databases. Starting from statistical analyses, a range of new tools are available, including metabolite set enrichment analysis, pathway and network visualization, pathway prediction, biochemical databases and text mining. Integrating these approaches into comprehensive and unbiased interpretations must carefully consider both caveats of the metabolomics dataset itself as well as the structure and properties of the biological study design. Special considerations need to be taken when adopting approaches from genomics for use in metabolomics. R and Python programming language are enabling an easier exchange of diverse tools to deploy integrated workflows. This review summarizes the key ideas and latest developments in regards to these approaches.

## Graphical abstract




Corresponding author: Fiehn, Oliver (ofiehn@ucdavis.edu).

## Introduction

Metabolomics aims to understand one of the fundamental questions in biology – how does metabolism interact with genetic and environmental factors? Studies reach from CRISPR/Cas9 mediated precise genetic changes [1] to nutritional [2] or environmental exposures [3], randomized clinical trials [4], [5] or efforts to enable the precision medicine [6]. Untargeted metabolite detection methods have let us realize that a substantial portion of the (human) metabolome is yet to be discovered [7]. Metabolomics datasets consist of thousands of compounds with up to 800 structurally identified metabolites [8]●●. Identified metabolites are chemically diverse and extend far beyond canonical energy or biopolymer biochemical pathways, enacting signaling functions as well as inhibitors or stress response systems. Capacity for obtaining metabolomics data is sufficient to routinely support general biomedical, preclinical, clinical and epidemiological studies [1,8–11]. A range of metabolomics datasets are publicly available with extensive biological metadata for facilitating investigations by other researchers [12] and the dbGaP database (https://www.ncbi.nlm.nih.gov/gap) study id ‘phs001334’ - Metabolomics of Coronary Heart Disease (CHD) in the Women’s Health Initiative (WHI).

However, as metabolomics datasets grow in size and complexity, it is becoming increasingly challenging to efficiently interpret changes in metabolite levels and determine their biological and clinical significance. Independent computational approaches for performing statistics, enrichment, visualization and contextualization need to be combined into integrated workflows that are tailored to specific study designs to extract comprehensive and meaningful information from the metabolomics datasets (Figure 1). Similar to the field of genomics research, metabolomics is moving towards building integrated workflows [13] ● [14] ●● [15]. This review provides an overview of recent approaches that can be included in such an integrated approach.

### What is the status of metabolomics data acquisition?

It is now well recognized that the chemical diversity of metabolites requires multiple assays to be combined to sufficiently cover the complexity of metabolomes [8,16]. Today, mass spectrometry is the most common tool used, mostly in combination with either gas- or liquid chromatography. Generally, high mass resolution instruments are used, with fast acquisition speed, great sensitivity and robust data acquisitions, including high chromatographic retention time stability. When internal standards are used, it has been shown that final result data sets can be harmonized and independent of the specific instrument being used, even for large series of samples [17]. Metabolomic datasets now extends to dozens of different chemical classes, including both endogenous and exogenous chemicals such as food and drug components [8]. New computational tools are being developed to standardize the data processing pipelines and to elucidate chemical structures for the detected unknown compounds in untargeted assays [18–21].

Metabolomics data has become a commodity. Metabolomics core and service centers have emerged in private and public sectors to allow routine purchase of high quality metabolomics datasets [8,22]. For example, in the United States, the National Institutes of Health have provided a major boost to the field by providing the seed funding to several

metabolomics service and research centers and by supporting a national metabolomics data repository (https://commonfund.nih.gov/metabolomics). These centers are developing and offering advanced assays and computational tools to cover metabolites beyond the core metabolic pathways [23]. With the growing scientific interest in using metabolomics in various study designs, these service centers will be instrumental in providing high-quality metabolomics datasets. However, the main challenge is to interpret these datasets by the investigators.

## What types of study designs are used in metabolomics?

Any metabolomic study starts with carefully defining the study designs to test specific metabolic questions. Study designs all come with specific strengths and limitations, and it is important to understand what can be confidently inferred from specific designs and matrices. For example, plasma samples from cross-sectional human cohorts are most suitable for finding diagnostic or exposure biomarkers, but less suitable for mechanistic interpretations on the basis of biochemical pathways. Nevertheless, multiple observational cohort studies have been used in metabolomics. Most suitable are longitudinal designs where cases are matched to control subjects to address major confounding criteria in nested case control designs. Observational case-control studies have two groups exposed and unexposed for which outcome events are counted and then associated with the metabolite levels. Typically, such study designs facilitate identifying risk factors for chronic diseases or aberrant metabolic phenotypes in tumor tissues, characterizing tumor sub-types and finding metabolite correlation modules. On the other hand, laboratory studies using animal models or cell cultures can be well well-controlled, but usually involve few samples per study group, and do not sufficiently reflect metabolic diversity under real life conditions (including variance in microbiomes).

## What types of statistics can be used in metabolomics?

Metabolomics statistical analyses is usually grouped into two categories depending on the interpretation level: univariate statistics and multivariate approaches. Statistical power is hard to be defined in metabolomics studies: even for univariate analyses, both effect sizes and within-group variance are usually not known beforehand. For multivariate statistics, approaches to power estimates have not been established. We are here making the case for using a third level of statistical analysis, using sets of variables.

**Univariate and multivariate statistics—**For univariate analyses, each metabolite is used separately as input for a statistical test. Inherently in univariate analysis, statistical independence is assumed for each variable (metabolite). Hence, statistical significance must be adjusted for multiple testing, especially if used for diagnostic purposes. Metabolome data usually follow non-Gaussian distributions and can be supposed to be with non-equal variance between test groups, requiring non-parametric significance tests. For epidemiological and clinical settings, regression models are used which can be adjusted for confounding variables such as age, gender and body mass index. Specially, for nested case control studies, conditional logistic regression models are used and for prospective cohort studies and clinical trials, cox proportional hazard models are used. In these regression models, effect size is reported as relative risk, odds ratio or hazard ration.

While univariate analysis is powerful for diagnostic questions, its major assumption of statistical independence of variables is simply untrue. Metabolite levels are not independent from each other, but are directly connected via myriads of enzymes and control steps. Hence, multivariate statistics explicitly exploits correlations between metabolites to obtain global metabolic phenotypes and to discriminate between groups of samples. Here, unsupervised exploratory data analyses (without using group assignments) or multi-variate regression and classification models are computed (including group labels in the analyses). Classic tools include the principal component analysis, hierarchical cluster analyses, support vector machines, random forest, partial least square-discriminant analyses and other tools. Classification models need data matrices to be divided into training and validation sets and show model's specificity and sensitivity as a receptor operative curve analysis. New approaches are emerging include deep neural networks [24], t-distributed stochastic neighbor embedding (t-sne) [25], lightgbm and xgboost which provide better speed and accuracy in training models or visualizing the global variance. Several reports have highlight that how much metabolomics data can improve the prediction accuracy for a diagnosis problem such as early diagnosis for lung cancer [26]●.

These univariate and multivariate level statistics are often calculated using R or SAS scripts but graphical user interface based tools have emerged for classical statistical tests [13–15]. These tools are useful for simple projects involving 2–3 groups. For complex studies, users still need to write R scripts to perform comprehensive statistical analysis.

**Metabolite set statistics—**While univariate analysis methods miss the systematic environment of metabolites and their inter-dependencies, multivariate methods oversimplify and do not consider biological relatedness. In between both approaches, metabolomics analyses can adopt ideas from genomic assessments, bridging statistical procedures with biological insights.

Set level statistics uses raw significance values from univariate statistics as input and apply these to sets of biologically connected metabolite groups. For each set of metabolites, an overall 'enrichment' statistics is then calculated to test whether these groups were altogether affected in a study. Hence, a very important first task in metabolite set enrichment statistics is to define the groups of metabolites that are biologically related to each other. Statistical outputs of metabolite enrichment patterns are easier to interpret in comparison to classic univariate- or multivariate statistics, as these groups are already attributed with biological functions. There are two categories of statistical tests for enrichment analysis, count based or distribution based [27]. Count based approach uses hypergeometric test or Fisher Exact tests whereas distribution based approach uses the Kolmogorov–Smirnov (KS) test. Distribution based test has the advantage that it does not need a univariate significance value cutoff; instead, it uses p-value distributions to calculate an overall set enrichment significance (Table1). Secondly, the KS test does not depend on the size of a background database for sampling assumptions [27]●●. Since metabolome data are ill-defined with respect to the complement of all possible metabolites that might be detected in a study (unlike genomics), enrichment statistics differ greatly if a hypergeometric (or Fisher Exact) test is used on a small database such as the KEGG ligand repository, or a large database, such as PubChem. In principle, any group of metabolites can be functionalized as set for calculating set

enrichments. Popular metabolite set definitions can be pathway maps, chemical classes [28] and metabolic modules that are derived from correlation networks or reaction networks.

Of particular interest in the scientific community are pathway analysis, because pathways directly lend functional roles to sets of individual metabolites [15]. Unfortunately, however, the very concept of metabolic pathways is not uniquely defined.

Three major problems undermine pathway analyses: (a) *Pathway databases for metabolites are incomplete.* Half of the detected metabolites in a typical metabolomics dataset do not have pathway annotations in existing biochemical databases as we here exemplify for a published study on non-obese diabetic mice [29] (Figure 2).

(b) *Metabolic pathways are manually defined and, hence, vary across different databases*. Some interpretation tools use KEGG pathway maps that include overlapping and intersecting pathways but largely disregard cellular compartmentalization. Other databases such as MetaCyc define pathways as uninterrupted linear sequence of enzymatic reactions, yielding an overall 2,526 pathways, many more than found in KEGG. (c) *Pathway enrichment statistics is ill-defined*. Many metabolites in metabolomic datasets are often member of different pathways, leading to difficulties in interpretation as well as in multiple testing problems. The main idea in pathway enrichment statistics [15] has been inspired by genomics: instead of relying only on univariate significance, molecules need to be grouped by biological relevance and tested if the group itself shows significant differences. However, unlike the number of genes, the number of metabolites in multicellular organisms is not known, partly because of enzyme ambiguity and certainly because the chemistry of life also involves nutrients and symbiotic relationships with the microbiome. The statistical tools currently used for pathway analyses are inaccurate, are not reproducible, subjective to interpretation bias, and lead to incorrect conclusions. Despite all these limitations, pathway maps are extensively used for metabolomics and integrated enrichment analysis of multi-omics datasets [30]. An alternative to pathway maps could be the organization of metabolites by ontology terms. For genomics and proteomics datasets, gene ontology (GO) terms are available, but metabolites have not been linked to the gene ontology terms yet.

Yet, a different type of biologically relevant ontology can be computed, the chemical similarity of metabolites themselves. Using relationship of chemical similarities assumes that few enzymes control or interconvert these compounds, even if the exact nature of these enzymes is not known. Chemical class ontologies are provided by the Medical Subject Heading ontology (MeSH) as well as by the European Bioinformatics Institute (ChEBI). Both ontologies have been proposed to be used for metabolite set enrichment analyses [31,32]. ClassyFire uses the ChEBI ontology as a reference database to predict classes for chemical compounds using a substructure finding method [33]. An alternative is to compute chemical similarities by comparing substructure fingerprints, as is often performed in pharmacology research [34] Chemical ontologies (and substructure similarity clustering) have two major advantages over pathway definitions for calculating enrichment statistics. First, almost every metabolite can be annotated with a chemical class, even if it is not annotated in enzyme-based pathway maps. Second, class definitions can be set such that each metabolite will belong to exactly one class only, yielding non-overlapping set

definitions for the enrichment computation. Unfortunately, neither ChEBI nor MeSH ontologies were sufficiently covering all detected metabolites in metabolomics datasets [35] (Figure 2). Recently, progress has been made by combining chemical class ontologies and chemical similarity mapping to annotate each metabolite with a chemical class. The approach and has been published as ChemRICH [35]●●, along with a web-based calculation tool (http://chemrich.fiehnlab.ucdavis.edu/). ChemRICH uses the KS-test for calculating enrichment statistics using the significance and directional changes of all identified metabolites (by SMILES and InChI keys) as input. ChemRICH yields named chemical clusters that are visualized by set enrichment significance and average set lipophilicity (Figure 3). ChemRICH plot highlights the significant chemical classes in a study to be interpreted in the context of class level biochemical processes. For example, an increase in triacylglycerols (TGs) and decrease in phosphophatidylcholines (PCs) lipids may indicate a remodeling of overall lipid metabolism towards storage lipids instead of membrane lipids, e.g. by increasing cell size and lipid droplets. While ChemRICH was shown to highlight new biological regulations for a previously published data set [35], it cannot directly be integrated with genomics or proteomics results.

Such multi-omic integrations would always need direct links between metabolites and enzymes. Approaches are being developed to use biochemical reaction network modules as set definitions [36], however they have the same two major bottlenecks outlined above (using the hypergeometric test and failing to cover all metabolites in biochemical reaction networks). Future research may integrate these approaches by overlaying chemical-structure focused sets with enzyme-derived reaction modules.

## How can we visualize metabolite relationships?

Metabolites are related to other metabolites through enzymatic reactions, chemical reactions, but they are also connected via structure similarities or through mathematical correlation of concentrations levels. Such relationships can be mapped by themselves or in combinations.

**Pathway maps—**Metabolite pathway diagrams are a traditional way to summarize and visually represent biochemical reactions [37] and usually comprise 10–20 metabolites per diagram. These maps are manually defined by researchers and thus follow different logics to define pathway boundaries. There is little consensus on the design of pathway map diagrams and what metabolites shall be included. Even for well-known maps such as the tricarboxylic acid cycle, different databases give different pathway maps [38]. There has been no effort to standardize pathway maps across various databases and generate a single consensus map.

As metabolomics datasets cover many metabolic pathways [39], tools to automatically create combined maps have been developed [40]. Particularly, integrated pathway maps are reported for metabolic flux analysis and tumor metabolism studies [41]. For example, glycolysis, the citric acid cycle and the pentose phosphate pathway are visualized together to visualize data and regulation of central metabolism [42]. Global maps [43,44] aim to provide a zoomable metabolic bird-eye’s view. However, metabolomics data sets usually contain metabolic end-products such as fatty acids or amino acids but not intermediates such as acetyl-CoA or oxaloacetate. Hence, global pathway maps show many metabolites that are

not found in metabolomics data sets, and vice versa, metabolomics data sets comprise many compounds that are not found in global biochemical pathway maps.

Pathway prediction algorithms map substrate/product pairs to enzymatic activities and then to genes [45,46]. Metabolism enumeration algorithms can be used to find novel reactions that are not yet catalogued in biochemical databases [47]. However, these algorithms have not been used yet to define new pathways or pathway maps. Pathway databases provide initial templates which can be customized by adding more information or by changing graph layouts [48]. Pathway diagrams should include genes, protein descriptors and transporter details to inform on tissue and species specificity. Such detailed and customizable maps may be particularly useful for visualizing results from studies with metabolomics, transcriptomics and proteomics datasets especially for tumor biopsies or tissue samples.

**Metabolic networks—**Some 3,000 human genes have been annotated with 1,200 enzymatic activities (E.C. numbers). Those enzymes can catalyze up to 4,000 known reactions which generate up to 2,500 human ‘endogenous’ metabolites [49]. Pathway maps are getting increasing complex the more metabolites are included, because most (central) metabolites participate in several or many biochemical reactions, depending on organelles, species and tissues. Alternatively, enzymatic reactions can be visualized as a metabolic network graph using a cluster layout algorithms. Graphs with, for example, 400 metabolites can still yield visually clear networks. For larger lists, network graphs can become very complex and dense, decreasing visual clarity and its utility to study metabolite relationships. A reaction network graph can be created with or without side-products or co-factors in biochemical reactions [50,51]. Ignoring the side-products improves the layout as several hub metabolites will be removed from the network graph.

An alternative to biochemical networks are using correlation structure of data sets or chemical (and mass spectral) similarity [52,53]. Such networks have the great advantage that metabolites do not need to be linked to specific enzymes, and (in principle) can also visualize unknowns. In addition, such approaches can be combined with enzyme reaction-based maps (KEGG RPair) in unified graphs through the MetaMapp tool [54], [55]. Using sparse partial correlation networks can even be used to bring unknown metabolites into reasonable biochemical modules [56]●.

## How to contextualize metabolites

Once a metabolite list or set list have been ranked by statistics, the next step is to find contexts in which those lists can be interpreted. It is beyond the capacity of an investigator to know all the relevant contexts and bioinformatics resource are instrumental here to assist the investigator. Two resources are useful for this purpose: biochemical databases and automated text mining. Using these resources, unbiased contexts can be identified that are useful for metabolomics interpretation.

**Querying databases—**Curated biochemical databases provides a wide range of information for metabolite which include reaction, enzymes, genes, regulatory metabolic genes, sources for exogenous metabolites, signaling properties, pathways, chemical classes,

known biological roles and diseases. Additionally, they can provide exact mass, lipophilicity, topological surface area, hydrogen bond donor and other chemical and physical properties.

**Key reaction databases are:** Brenda, which provide reactions and signaling properties of metabolites. It is the largest database of curated biochemical reactions and ligands with almost 120K mapped to it. MeSH and ChEBI, which are major curated ontologies for metabolites. MeSH is supported by national library of medicine and has mapped 120,000 compounds to almost 3000 chemical categories. Pathway databases are KEGG, Reactome, MetaCyc, Wikipathways and SMPDB. Notable pathway repositories are the NCBI BioSystems database and ConsensusPathDB which provides single point access to pathway maps from a range of databases. HMDB aims to catalogue all the metabolites that are relevant for human [57] and provide extracted information from biomedical literature. These databases provide online interfaces but automated queries using web-APIs allows running a large number of queries using a programming language [58].

**Text mining—**Manually curated databases have high quality content extracted from literature database. However, the biomedical literature volume is growing exponentially, and the slow manual curation process may not be able to process all the relevant publications. PubMed lists 30 million total references and almost 7.2 million reports are related to metabolism. Manually extracting information from that many papers is not feasible. Therefore, we need to use computational text mining [59] and analytics [58] approaches to extract relevant information from the literature. FACTA provide MeSH aggregations to pin-point associations of metabolite with biomedical terms [60]. Table 2 shows such aggregation for hydroxy-proline. MeSH term aggregations determine the frequency how often specific terms are associated with metabolites. Such associations can be tested by statistical significance, for example, to link metabolites with diseases. This approach has been used to build the MetDisease database [61]. The NutriChem database [62] has been developed using a similar approach to find plant and diet related compounds from PubMed. Metabolomics text mining was used to extract information on all literature-known compounds in yeast [63], and to complement pathway reconstructions through reports on product/substrate pairs [64]. However, there has been little progress in using automated text mining approaches for the complement of metabolomics data sets, with the sole exception of PolySearch [65] ●●. Development of automated text mining tools using text analytics database tools such as ElasticSearch or Apache Lucene will be the next innovative breakthrough in metabolomics data processing and interpretational approaches.

## The path forward

Computational approaches for metabolomics interpretation need to keep up with fast evolving metabolomics datasets generated by service centers. Tools need to improve enrichment analyses, building customizable and detailed pathway diagrams and comprehensively linking metabolites to biomedical contexts. Use of text analytics approaches such as ElasticSearch may find biomedical contexts for interpreting metabolomic outcomes against the overall scientific literature. Unknown metabolites need to be mapped to most similar known compounds and included into network modules based enrichment analysis methods. These approaches need to be integrated into workflows that

are tailored towards specific study designs, using R or Python programming languages for re-usability and testing.

## Acknowledgments

This work was funded by the grant NIH DK097154.

## References

1. Sperber H, Mathieu J, Wang Y, Ferreccio A, Hesson J, Xu Z, Fischer KA, Devi A, Detraux D, Gu H, et al. The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. Nat Cell Biol. 2015; 17:1523–1535. [PubMed: 26571212]
2. Kieffer DA, Piccolo BD, Marco ML, Kim EB, Goodson ML, Keenan MJ, Dunn TN, Knudsen KE, Martin RJ, Adams SH. Mice Fed a High-Fat Diet Supplemented with Resistant Starch Display Marked Shifts in the Liver Metabolome Concurrent with Altered Gut Bacteria. J Nutr. 2016; 146:2476–2490. [PubMed: 27807042]
3. Barupal DK, Pinkerton KE, Hood C, Kind T, Fiehn O. Environmental Tobacco Smoke Alters Metabolic Systems in Adult Rats. Chem Res Toxicol. 2016; 29:1818–1827. [PubMed: 27788581]
4. Wang DD, Toledo E, Hruby A, Rosner BA, Willett WC, Sun Q, Razquin C, Zheng Y, Ruiz-Canela M, Guasch-Ferre M, et al. Plasma Ceramides, Mediterranean Diet, and Incident Cardiovascular Disease in the PREDIMED Trial (Prevencion con Dieta Mediterranea). Circulation. 2017; 135:2028–2040. [PubMed: 28280233]
5. Tenori L, Oakman C, Morris PG, Gralka E, Turner N, Cappadona S, Fornier M, Hudis C, Norton L, Luchinat C, et al. Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study. Mol Oncol. 2015; 9:128–139. [PubMed: 25151299]
6. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; 372:793–795. [PubMed: 25635347]
7. Showalter MR, Cajka T, Fiehn O. Epimetabolites: discovering metabolism beyond building and burning. Curr Opin Chem Biol. 2017; 36:70–76. [PubMed: 28213207]
8●●. Karl JP, Margolis LM, Murphy NE, Carrigan CT, Castellani JW, Madslien EH, Teien HK, Martini S, Montain SJ, Pasiakos SM. Military training elicits marked increases in plasma metabolomic signatures of energy metabolism, lipolysis, fatty acid oxidation, and ketogenesis. Physiol Rep. 2017; 5
9. Huang J, Mondul AM, Weinstein SJ, Koutros S, Derkach A, Karoly E, Sampson JN, Moore SC, Berndt SI, Albanes D. Serum metabolomic profiling of prostate cancer risk in the prostate, lung, colorectal, and ovarian cancer screening trial. Br J Cancer. 2016; 115:1087–1095. [PubMed: 27673363]
10. Hakimi AA, Reznik E, Lee CH, Creighton CJ, Brannon AR, Luna A, Aksoy BA, Liu EM, Shen R, Lee W, et al. An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. Cancer Cell. 2016; 29:104–116. [PubMed: 26766592]
11. Miller DB, Ghio AJ, Karoly ED, Bell LN, Snow SJ, Madden MC, Soukup J, Cascio WE, Gilmour MI, Kodavanti UP. Ozone Exposure Increases Circulating Stress Hormones and Lipid Metabolites in Humans. Am J Respir Crit Care Med. 2016; 193:1382–1391. [PubMed: 26745856]
12. St John-Williams L, Blach C, Toledo JB, Rotroff DM, Kim S, Klavins K, Baillie R, Han X, Mahmoudiandehkordi S, Jack J, et al. Targeted metabolomics and medication classification data from participants in the ADNI1 cohort. Sci Data. 2017; 4:170140. [PubMed: 29039849]
13●. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017; 13:e1005752. [PubMed: 29099853]
14●●. Wanichthanarak K, Fan S, Grapov D, Barupal DK, Fiehn O. Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration. PLoS One. 2017; 12:e0171046. [PubMed: 28141874]
15. Xia JG, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Research. 2015; 43:W251–W257. [PubMed: 25897128]

16. Guo L, Milburn MV, Ryals JA, Lonergan SC, Mitchell MW, Wulff JE, Alexander DC, Evans AM, Bridgewater B, Miller L, et al. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. Proc Natl Acad Sci U S A. 2015; 112:E4901–4910. [PubMed: 26283345]

17. Cajka T, Smilowitz JT, Fiehn O. Validating Quantitative Untargeted Lipidomics Across Nine Liquid Chromatography-High-Resolution Mass Spectrometry Platforms. Anal Chem. 2017; 89:12360–12368. [PubMed: 29064229]

18. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015; 12:523–526. [PubMed: 25938372]

19. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Anal Chem. 2016; 88:7946–7958. [PubMed: 27419259]

20. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta SS, Wohlgemuth G, Barupal DK, Showalter MR, Arita M, et al. Identification of small molecules using accurate mass MS/MS search. Mass Spectrom Rev. 2017

21. Moorthy AS, Wallace WE, Kearsley AJ, Tchekhovskoi DV, Stein SE. Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification. Anal Chem. 2017

22. Zheng H, Powell JE, Steele MI, Dietrich C, Moran NA. Honeybee gut microbiota promotes host weight gain via bacterial metabolism and hormonal signaling. Proc Natl Acad Sci U S A. 2017; 114:4775–4780. [PubMed: 28420790]

23. Lai Z, Kind T, Fiehn O. Using Accurate Mass Gas Chromatography-Mass Spectrometry with the MINE Database for Epimetabolite Annotation. Anal Chem. 2017; 89:10171–10180. [PubMed: 28876899]

24. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, Ostrovskiy A, Cantor C, Vijg J, Zhavoronkov A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging (Albany NY). 2016; 8:1021–1033. [PubMed: 27191382]

25. Abdelmoula WM, Balluff B, Englert S, Dijkstra J, Reinders MJ, Walch A, McDonnell LA, Lelieveldt BP. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proc Natl Acad Sci U S A. 2016; 113:12244–12249. [PubMed: 27791011]

26●. Wikoff WR, Hanash S, DeFelice B, Miyamoto S, Barnett M, Zhao Y, Goodman G, Feng Z, Gandara D, Fiehn O, et al. Diacetylspermine Is a Novel Prediagnostic Serum Biomarker for Non-Small-Cell Lung Cancer and Has Additive Performance With Pro-Surfactant Protein B. J Clin Oncol. 2015; 33:3880–3886. [PubMed: 26282655]

27●●. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nat Rev Genet. 2016; 17:353–364. [PubMed: 27070863]

28. Lopez-Ibanez J, Pazos F, Chagoyen M. MBROLE 2.0-functional enrichment of chemical compounds. Nucleic Acids Res. 2016; 44:W201–204. [PubMed: 27084944]

29. Fahrmann J, Grapov D, Yang J, Hammock B, Fiehn O, Bell GI, Hara M. Systemic alterations in the metabolome of diabetic NOD mice delineate increased oxidative stress accompanied by reduced inflammation and hypertriglyceremia. Am J Physiol Endocrinol Metab. 2015; 308:E978–989. [PubMed: 25852003]

30. Biancur DE, Paulo JA, Malachowska B, Quiles Del Rey M, Sousa CM, Wang X, Sohn ASW, Chu GC, Gygi SP, Harper JW, et al. Compensatory metabolic networks in pancreatic cancers upon perturbation of glutamine metabolism. Nat Commun. 2017; 8:15965. [PubMed: 28671190]

31. Moreno P, Beisken S, Harsha B, Muthukrishnan V, Tudose I, Dekker A, Dornfeldt S, Taruttis F, Grosse I, Hastings J, et al. BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. BMC Bioinformatics. 2015; 16:56. [PubMed: 25879798]

32. Tsuyuzaki K, Morota G, Ishii M, Nakazato T, Miyazaki S, Nikaido I. MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. BMC Bioinformatics. 2015; 16:45. [PubMed: 25887539]

33. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016; 8:61. [PubMed: 27867422]
34. Nickel J, Gohlke BO, Erehman J, Banerjee P, Rong WW, Goede A, Dunkel M, Preissner R. SuperPred: update on drug classification and target prediction. Nucleic Acids Res. 2014; 42:W26–31. [PubMed: 24878925]
35●●. Barupal DK, Fiehn O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep. 2017; 7:14567. [PubMed: 29109515]
36. Picart-Armada S, Fernandez-Albert F, Vinaixa M, Rodriguez MA, Aivio S, Stracker TH, Yanes O, Perera-Lluna A. Null diffusion-based enrichment for metabolomics data. PLoS One. 2017; 12:e0189012. [PubMed: 29211807]
37. Murray P, McGee F, Forbes AG. A taxonomy of visualization tasks for the analysis of biological pathway data. BMC Bioinformatics. 2017; 18:21. [PubMed: 28251869]
38. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013; 14:112. [PubMed: 23530693]
39. Liesenfeld DB, Grapov D, Fahrmann JF, Salou M, Scherer D, Toth R, Habermann N, Bohm J, Schrotz-King P, Gigic B, et al. Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. Am J Clin Nutr. 2015; 102:433–443. [PubMed: 26156741]
40. Paley S, O'Maille PE, Weaver D, Karp PD. Pathway collages: personalized multi-pathway diagrams. BMC Bioinformatics. 2016; 17:529. [PubMed: 27964719]
41. Park JO, Rubin SA, Xu YF, Amador-Noguez D, Fan J, Shlomi T, Rabinowitz JD. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. Nat Chem Biol. 2016; 12:482–489. [PubMed: 27159581]
42. Luo W, Pant G, Bhavnasi YK, Blanchard SG Jr, Brouwer C. Pathview Web: user friendly pathway visualization and data integration. Nucleic Acids Res. 2017
43. Sidiropoulos K, Viteri G, Sevilla C, Jupe S, Webber M, Orlic-Milacic M, Jassal B, May B, Shamovsky V, Duenas C, et al. Reactome enhanced pathway visualization. Bioinformatics. 2017; 33:3461–3467. [PubMed: 29077811]
44. Kelley JJ, Maor S, Kim MK, Lane A, Lun DS. MOST-visualization: software for producing automated textbook-style maps of genome-scale metabolic networks. Bioinformatics. 2017; 33:2596–2597. [PubMed: 28430868]
45. Tabei Y, Yamanishi Y, Kotera M. Simultaneous prediction of enzyme orthologs from chemical transformation patterns for de novo metabolic pathway reconstruction. Bioinformatics. 2016; 32:i278–i287. [PubMed: 27307627]
46. Pertusi DA, Stine AE, Broadbelt LJ, Tyo KE. Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics. 2015; 31:1016–1024. [PubMed: 25417203]
47. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KE, Henry CS. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Cheminform. 2015; 7:44. [PubMed: 26322134]
48. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, Evelo CT. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015; 11:e1004085. [PubMed: 25706687]
49. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013; 31:419–425. [PubMed: 23455439]
50. Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot MF, Jourdan F. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Res. 2010; 38:W132–137. [PubMed: 20444866]
51. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, Laudanna C, Sartor MA, Stringer KA, Jagadish HV, Burant C, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics. 2012; 28:373–380. [PubMed: 22135418]

52. Quinn RA, Nothias LF, Vining O, Meehan M, Esquenazi E, Dorrestein PC. Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. Trends Pharmacol Sci. 2017; 38:143–154. [PubMed: 27842887]

53. Skogerson K, Wohlgemuth G, Barupal DK, Fiehn O. The volatile compound BinBase mass spectral database. BMC Bioinformatics. 2011; 12:321. [PubMed: 21816034]

54. Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton KE, Fiehn O. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinformatics. 2012; 13:99. [PubMed: 22591066]

55. Grapov D, Wanichthanarak K, Fiehn O. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. Bioinformatics. 2015; 31:2757–2760. [PubMed: 25847005]

56●. Basu S, Duren W, Evans CR, Burant CF, Michailidis G, Karnovsky A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. Bioinformatics. 2017; 33:1545–1553. [PubMed: 28137712]

57. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2017

58. Guha N, Guyton KZ, Loomis D, Barupal DK. Prioritizing Chemicals for Risk Assessment Using Chemoinformatics: Examples from the IARC Monographs on Pesticides. Environ Health Perspect. 2016; 124:1823–1829. [PubMed: 27164621]

59. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012; 13:829–839. [PubMed: 23150036]

60. Tsuruoka Y, Tsujii J, Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics. 2008; 24:2559–2560. [PubMed: 18772154]

61. Duren W, Weymouth T, Hull T, Omenn GS, Athey B, Burant C, Karnovsky A. MetDisease–connecting metabolites to diseases via literature. Bioinformatics. 2014; 30:2239–2241. [PubMed: 24713438]

62. Jensen K, Panagiotou G, Kouskoumvekaki I. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. Nucleic Acids Res. 2015; 43:D940–945. [PubMed: 25106869]

63. Nobata C, Dobson PD, Iqbal SA, Mendes P, Tsujii Ji, Kell DB, Ananiadou S. Mining metabolites: extracting the yeast metabolome from the literature. Metabolomics. 2011; 7:94–101. [PubMed: 21687783]

64. Czarnecki J, Nobeli I, Smith AM, Shepherd AJ. A text-mining system for extracting metabolic reactions from full-text articles. BMC bioinformatics. 2012; 13:172. [PubMed: 22823282]

65●●. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res. 2015; 43:W535–542. [PubMed: 25925572]

## Highlights

- Statistics has become a commodity and serves as entry into data interpretation.
- Metabolic pathway databases are incomplete and disagree on pathway definitions.
- Metabolite set enrichment analysis can be independent of background databases.
- Network graph based approaches can include unknown metabolites into interpretation.
- Automated text mining extends the interpretation to include biological contexts.

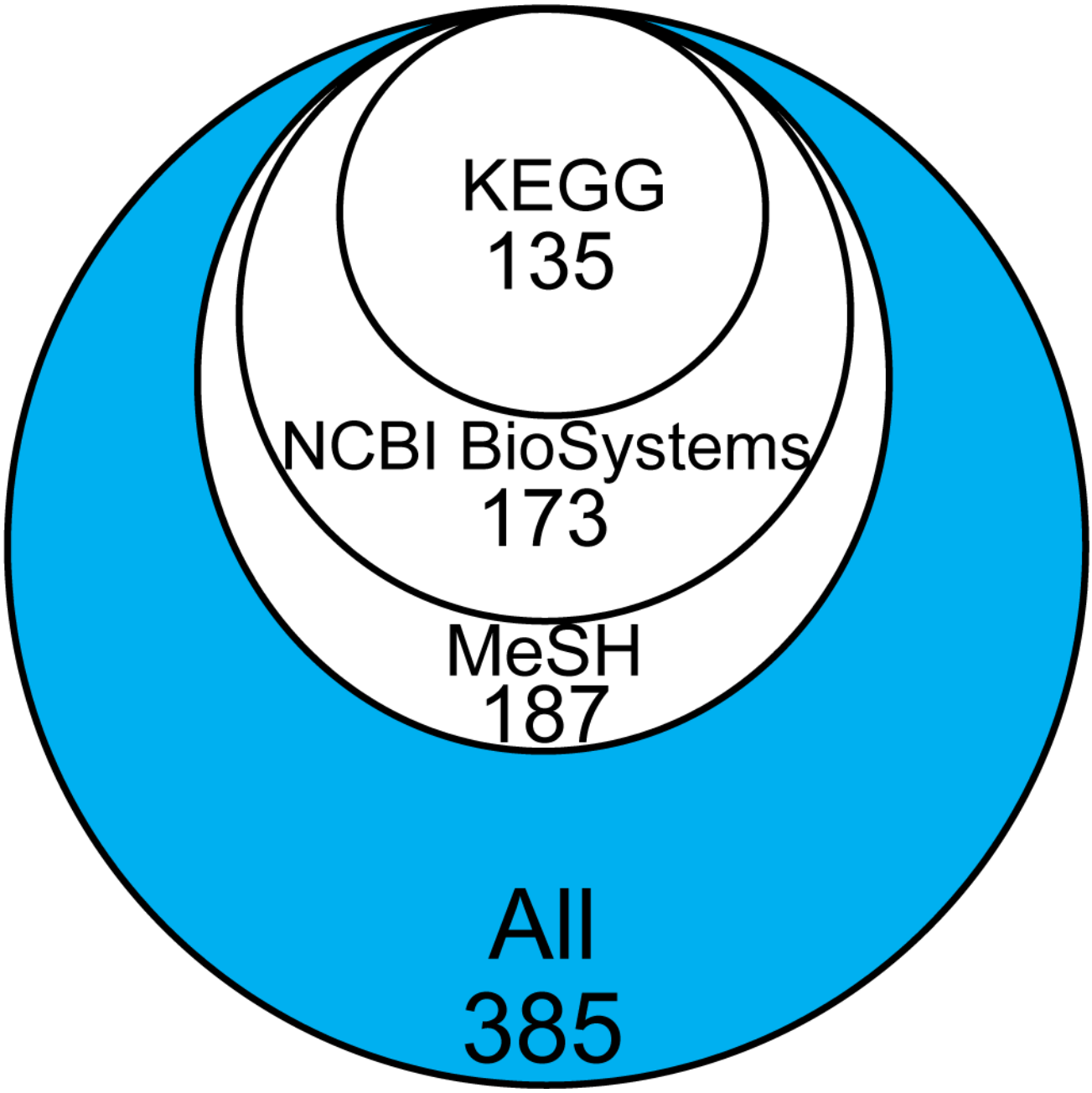



**Figure 1.**
Metabolomics interpretation approaches can be combined into study-design specific workflows to provide a comprehensive interpretation.

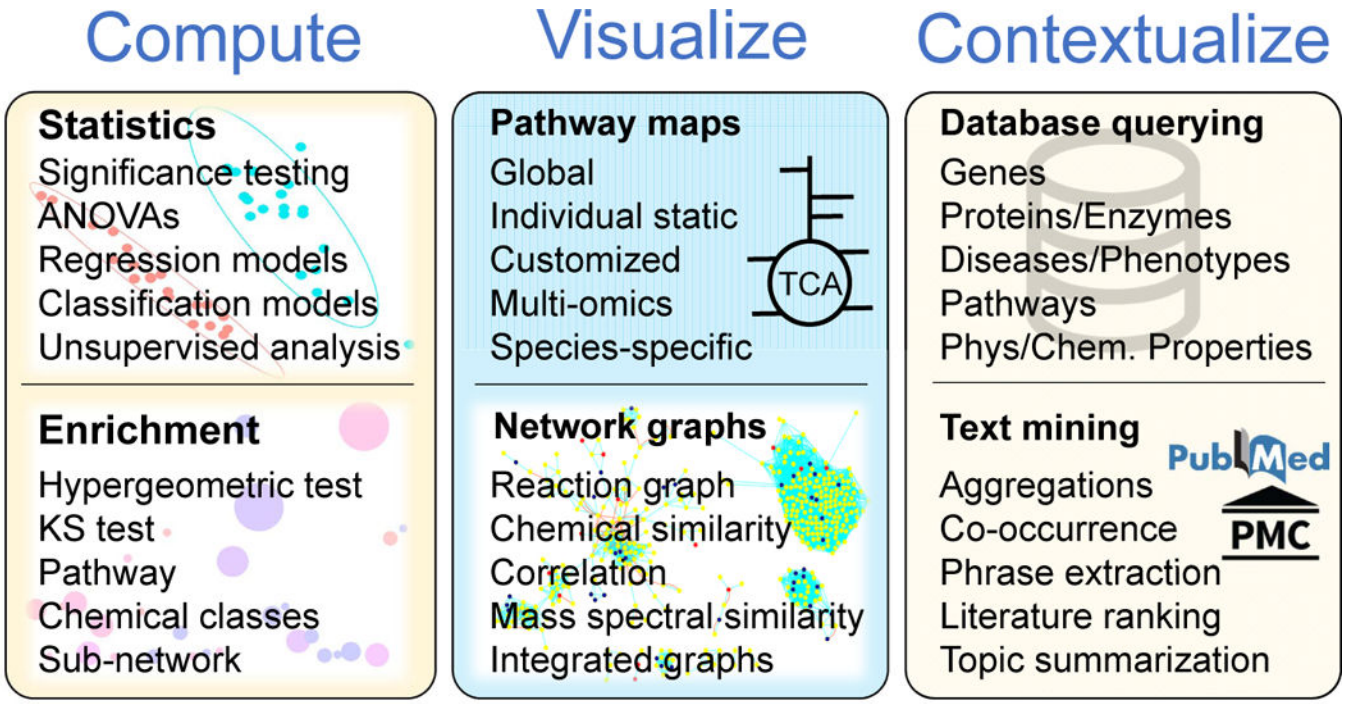



**Figure 2.**

Biochemical and ontology databases lack entries for the 385 metabolites identified in non-obese diabetic mice [30]. Figure is adopted from [35].

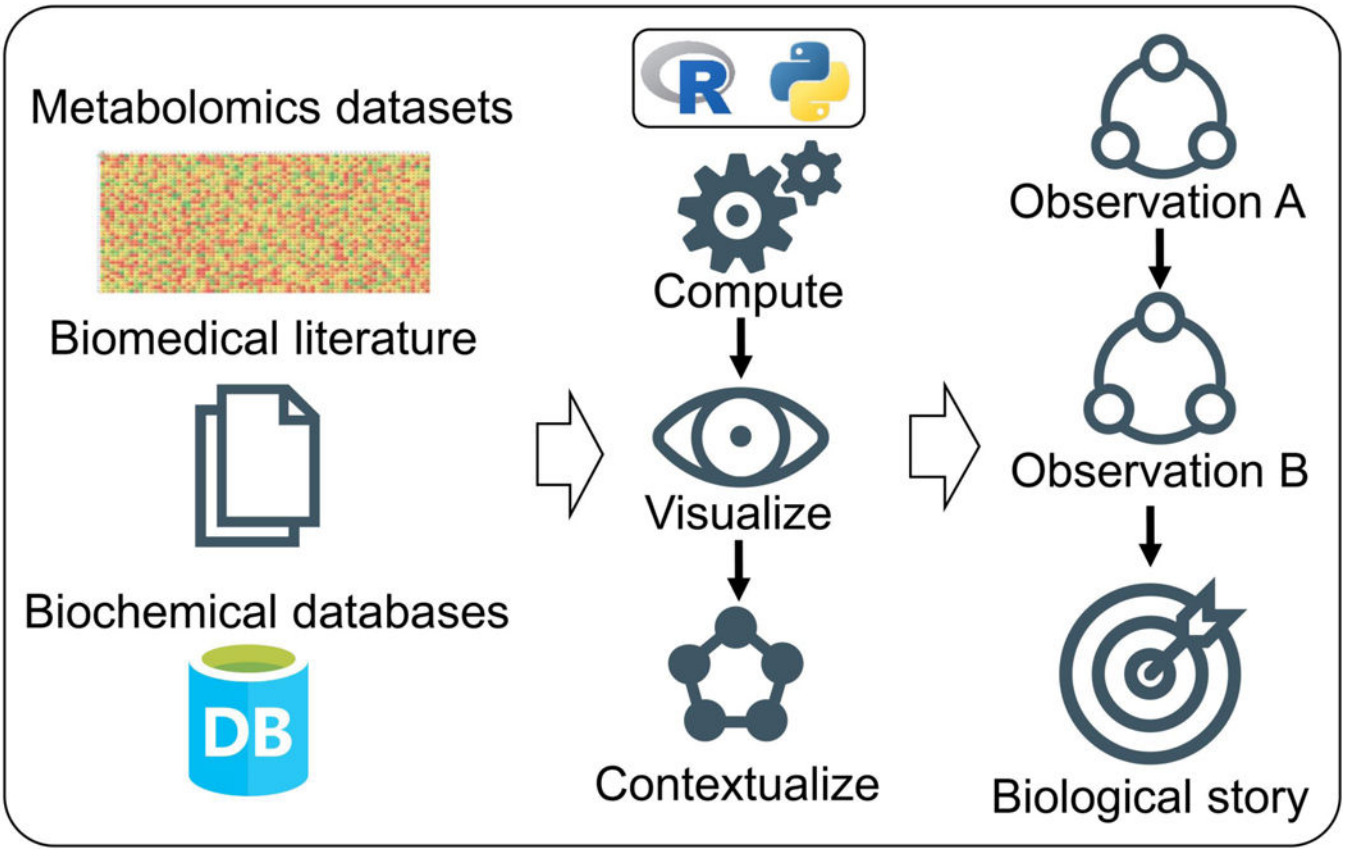



**Figure 3.**

ChemRICH impact plot of chemical similarity enrichment analysis in non-obese diabetic mice [30]. Color indicates direction of change for most of the compounds within a class : red is increased, blue is decreased. Size of cluster indicates the number of metabolites within the class. X-axis shows the cluster order on the Tanimoto similarity tree. The tree order also correlates with average lipophilicity of classes so polar classes are always shown on the right side of the plot. Y-axis shows the negative log of adjusted p-values so significantly important classes are shown at the top of the plot. Figure is adopted from [35].

**Table 1**

Classical statistical tests for enrichment depends on background database and p-value cutoffs.

| Parameter | Fisher-exact | Hypergeometric | Bionomial | Kolmogorov–Smirnov |
|---|---|---|---|---|
| Background Database | Yes | Yes | No | No |
| P-value cutoff | Yes | Yes | Yes | No |

## Table 2

MeSH term aggregation for Hydroxyproline. Numbers show the citation count in PubMed database.

| Medical Subject Headings | | | |
|---|---|---|---|
| **Processes** | **Enzymes** | **Diseases** | **Chemical and drugs** |
| Time Factors:88 | Superoxide Dismutase:37 | Pulmonary Fibrosis:336 | Hydroxyproline:633 |
| Organ Size:80 | Procollagen-Proline Dioxygenase:21 | Silicosis:29 | Bleomycin:233 |
| Body Weight:61 | Peroxidase:15 | Lung Diseases:21 | Collagen:208 |
| Dose-Response | L-Lactate | Pneumonia:21 | Silicon Dioxide:50 |
| Relationship, Drug:40 | Dehydrogenase:14 | Lung Injury:16 | RNA, Messenger:48 |
| Cell Count:38 | Matrix Metalloproteinase 9:12 | Lung Neoplasms:14 | Transforming Growth Factor beta1:48 |
| Cell Division:26 | | Pulmonary Emphysema:12 | |