OXFORD

Briefings in Bioinformatics, 2024, **25(6)**, bbae498
https://doi.org/10.1093/bib/bbae498
Review

Current approaches and outstanding challenges of functional annotation of metabolites: a comprehensive review

Quang-Huy Nguyen[1], Ha Nguyen[1], Edwin C. Oh[2], Tin Nguyen[1,*]

[1]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, United States
[2]Department of Internal Medicine, UNLV School of Medicine, University of Nevada, Las Vegas, NV 89154, United States
*Corresponding author: Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA. Tel.: +1 334 844 4359; E-mail: tinn@auburn.edu

Abstract

Metabolite profiling is a powerful approach for the clinical diagnosis of complex diseases, ranging from cardiometabolic diseases, cancer, and cognitive disorders to respiratory pathologies and conditions that involve dysregulated metabolism. Because of the importance of systems-level interpretation, many methods have been developed to identify biologically significant pathways using metabolomics data. In this review, we first describe a complete metabolomics workflow (sample preparation, data acquisition, pre-processing, downstream analysis, etc.). We then comprehensively review 24 approaches capable of performing functional analysis, including those that combine metabolomics data with other types of data to investigate the disease-relevant changes at multiple omics layers. We discuss their availability, implementation, capability for pre-processing and quality control, supported omics types, embedded databases, pathway analysis methodologies, and integration techniques. We also provide a rating and evaluation of each software, focusing on their key technique, software accessibility, documentation, and user-friendliness. Following our guideline, life scientists can easily choose a suitable method depending on method rating, available data, input format, and method category. More importantly, we highlight outstanding challenges and potential solutions that need to be addressed by future research. To further assist users in executing the reviewed methods, we provide wrappers of the software packages at https://github.com/tinnlab/metabolite-pathway-review-docker.

Keywords: liquid chromatography; mass spectrometry; metabolomics; functional analysis; metabolic pathways

Introduction

Metabolomics is a powerful approach for identifying predictive biomarkers of diseases, including obesity [1], diabetes [2, 3], cardiovascular disease [4, 5]) to cancer [6, 7], cognitive disorders [8], respiratory pathologies [9, 10], and other conditions that involve dysregulated metabolism [11–13]. There are two approaches to metabolite profiling: targeted and untargeted metabolomics (also called shotgun, global, or nontargeted). The former focuses on the quantitative measurement of a predefined set of known metabolites, while the latter aims to comprehensively analyze as many metabolites as possible, including unknown compounds, without prior selection. Recent technologies include nuclear magnetic resonance, and mass spectrometry (MS) with liquid chromatography (LC) or gas chromatography [14–16]. Among them, LC-MS has gained popularity due to additional detection of nonvolatile compounds, the resolution of individual chemical components into distinct peaks, and the ability to resolve isobaric compounds while minimizing signal suppression [14].

Regardless of technologies, a comparative study (disease versus healthy, treated versus control, etc.) often yields a set of metabolites or spectral features that are differentially expressed (DE) between the phenotypes. Although these lists of DE metabolites are important in identifying potential biomarkers, they alone fail to reveal the underlying mechanism. The challenge is to move beyond differential expression analysis to determine the biological and physiological roles of metabolites at the systems level [11]. For that purpose, researchers have created various knowledge bases that group metabolites, genes, and gene products into functional modules and interaction networks, including KEGG [17], HMDB [18], BioCyc [19], Reactome [20], PubChem [21], etc.

Concurrently, computational methods have been developed to identify impacted pathways or perturbed functional modules. These include basic and advanced levels [22, 23]. The former involves mapping metabolites to pathways and visualization, while the latter utilizes sophisticated enrichment and statistical analysis. In this review, we focus on advanced methods that (1) can perform functional analysis, and (2) have software that is maintained since 2018. Among these, the earliest approaches use Over-Representation Analysis (ORA) [24, 25] to identify functional modules that have DE entities (genes/metabolites) that exhibit greater variations between different conditions than expected by chance. The drawbacks of this type of approach include that (i) it only considers the number of DE entities and ignores the magnitude of the actual changes; (ii) it assumes that genes and

Received: May 22, 2024. **Revised:** September 3, 2024. **Accepted:** October 2, 2024

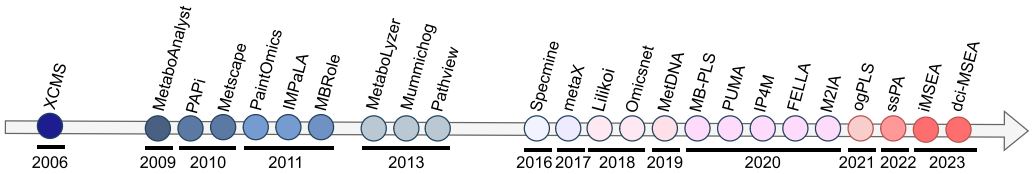

Figure 1. Chronological timeline of the 24 surveyed tools for functional analysis using either metabolomics data or integration with other multi-omics data. The timeline spans 2006 to 2023, highlighting the development and improvement of functional analysis methods over the past 17 years.

molecules are independent, which they are not; and (iii) it ignores the interactions between various modules. Functional Class Scoring (FCS) approaches [26–28] have been developed to address some of the issues raised by ORA. The main improvement of FCS is the observation that small, yet coordinated changes in expression of functionally related entities can have a significant impact on pathways. Topology-based Pathway Analysis (TPA) approaches leverage the topology of pathways and the interactions among omics features to more accurately represent the underlying biological phenomena [29, 30].

This review aims to provide a broad overview of the current state of functional annotation approaches designed for metabolomics data. Figure 1 shows the 24 methods included in our review. Our objectives are to (1) describe the main categories of functional analysis approaches, (2) discuss methodological foundations, (3) highlight strengths and limitations, and (4) identify the outstanding challenges and future directions in the field. Existing articles review tools for metabolomics analysis but do not provide an in-depth investigation into the functional analysis [22, 31–35]. For example, Stanstrup *et al.* [34] outline the available resources for preprocessing, statistical analysis, and the visual representation of data that assists in identifying metabolic changes in various biological settings. In contrast, we provide a comprehensive review of prominent methods capable of performing functional analysis.

The manuscript is organized as follows. Section Availability and implementation describes the distinct characteristics of the 24 approaches: availability, platform, data processing, quality control, input types, databases, and supported analyses. Section Complete metabolomics workflow summarizes the complete workflow of metabolomics studies, starting from sample extraction, data acquisition, quality control to statistical analysis and pathway-level analysis. Section Annotation and omics-level analysis details downstream analyses at the metabolite and omics level. Section Pathway-level analysis categorizes the methods into distinct categories based on their techniques and discusses their methodologies. Section Summary and practical guideline compares and contrasts the surveyed tools and provides readers with specific guidelines for selecting appropriate methods in certain situations. Section Outstanding challenges and solutions highlights challenges that remain unsolved in the field. We also provide a summary of every single method in Supplementary Note.

Availability and implementation

Table 1 shows the availability of the reviewed methods: reference, software link, publication year, latest update, number of citations, and platform. The prevalence of web-based and R-based approaches becomes evident when considering the most commonly selected platforms for metabolic pathway analysis. Figure 2 shows the essential information related to the functionality and methodogologies of the approaches. The second column shows the methods that support performing data preprocessing and quality control for three-dimensional files. The third column indicates whether they support global/untargeted metabolomics data (versus targeted metabolomics). If a method supports global metabolomics, it can be one of the two following cases or both: (i) the method integrates Mummichog [36] that can perform functional analysis without annotation of metabolites, and (ii) the method performs metabolite annotation from spectral features before performing statistical analysis (e.g. marker detection) and functional analysis. The next five columns list the supported input data: genomics (G), transcriptomics (T), epigenomics (E), proteomics (P), and metabolomics (M). The next column shows the pathway databases embedded in each software. The remaining columns show the functionalities implemented in each software: (1) metabolite annotation, (2) sample clustering, (3) differential analysis or marker identification, (4) functional analysis from spectral features (F2P: feature to pathway), (5) functional analysis from annotated metabolites (M2P: metabolite to pathway), (6) network construction, or (7)meta-analysis (multi-cohort analysis).

We divide the tools into three categories based on their methodologies: (i) ORA, (ii) FCS, and (iii) TPA. There are eight tools that support metabolite annotations and 14 tools that allow users to perform differential analysis and marker detection. Among the eight tools that support untargeted metabolomics, only Mummichog, XCMS, and MetaboAnalyst perform functional analysis without upfront putative identification of metabolites. Mummichog was the first approach that employed such feature-to-pathway strategy (F2P), which was then followed by XCMS and MetaboAnalyst. XCMS also supports meta-analysis at the marker level (metabolites) whereas MetaboAnalyst supports meta-analysis at both marker and pathway levels. MetaboAnalyst stands out as a comprehensive implementation for metabolomics data analysis, encompassing various data input types, data processing methods, and statistical approaches. Notably, it is the only software that allows users to analyze genomics data.

Complete metabolomics workflow

Figure 3 provides a high-level workflow for metabolomics studies. Researchers start a study by defining the biological problem of interest and forming research hypothesis [60]. There are four main steps for the detection and analysis of metabolites: (A) sample preparation and extraction, (B) data acquisition using LC-MS instruments, (C) data processing and quality control, and (D) downstream analysis.

In sample preparation and extraction (step 1, Fig. 3(A), researchers seek to collect relevant biological specimens from research participants. In data acquisition (step 2, Fig. 3(B), researchers generate metabolomics data using high-resolution LC-MS instruments. In data processing and quality control (step 3, Fig. 3(C), researchers process the raw data and perform quality control using the 3D files in a vendor-neutral format. In downstream analysis (step 4, Fig. 3(D), users can perform

Table 1. Availability of 24 pathway analysis approaches

Method	Software Link	Pub. Year	Latest Update	Citation	Platform
Over-Representation Analysis (ORA)					
XCMS [37]	https://xcmsonline.scripps.edu/	2006	2018	5,063	🌐, 💻: R
PaintOmics [38, 39]	https://www.paintomics.org/	2011	2022	176	🌐
MBRole [40]	http://csbg.cnb.csic.es/mbrole3/	2011	2023	162	🌐
MetaboLyzer [41]	https://sites.google.com/a/georgetown.edu/fornace-lab-informatics/home/metabolyzer/	2013	2023	107	💻: Python
Mummichog [36]	https://github.com/shuzhao-li/mummichog	2013	2021	852	💻: Python
MetDNA [42]	http://metdna.zhulab.cn/	2019	2022	252	🌐
Functional Class Scoring (FCS)					
MetaboAnalyst [43]	https://www.metaboanalyst.ca/	2009	2023	2,870	🌐, 💻: R
PAPi [28]	https://rdrr.io/bioc/PAPi/	2010	2019	137	💻: R
IMPaLA [30]	http://impala.molgen.mpg.de/	2011	2021	417	🌐, 💻: Python
Pathview [44]	https://pathview.uncc.edu/guest-home	2013	2023	1,794	🌐, 💻: R
specmine [45]	https://webspecmine.bio.di.uminho.pt/	2016	2019	58	🌐, 💻: R
metaX [46]	https://github.com/wenbostar/metaX/	2017	2018	554	💻: R
Lilikoi [47]	https://cran.r-project.org/web/packages/lilikoi/	2018	2022	33	💻 R
MB-PLS [48]	https://github.com/jydong2018/metabolomics/	2020	2020	13	💻:MATLAB
PUMA [49]	https://github.com/HassounLab/PUMA/	2020	2020	10	💻: Python
IP4M [50]	https://github.com/IP4M/	2020	2020	39	💻: ☕,PERL
M2IA [51]	https://m2ia.met-bioinformatics.cn/	2020	2020	61	🌐
ogPLS [52]	https://github.com/jydong2018/ogPLS/	2021	2021	11	💻:MATLAB
ssPA [53]	https://pypi.org/project/sspa/	2022	2022	15	💻: Python
Topology-based Pathway Analysis (TPA)					
Metscape [54, 55]	http://metscape.med.umich.edu/	2010	2018	436	💻: ☕
Omicsnet [56]	https://www.omicsnet.ca/	2018	2022	163	🌐
FELLA [57]	https://github.com/b2slab/FELLA/	2020	2020	84	💻: R
iMSEA [58]	https://github.com/BioNet-XMU/iMSEA/	2023	2023	5	💻:MATLAB
dci-MSEA [59]	https://github.com/BioNet-XMU/dci-MSEA/	2023	2023	1	💻: Python

Abbreviation: 🌐: web application, 💻: standalone software, R: R, Python: Python, ☕: Java or Javascript, MATLAB: Matlab, PERL: Perl. Software with 🌐 are those available as web-based platform while those with 💻 are available as standalone package. For standalone packages, we also specify the programming language used to develop the software. Citations were retrieved from Google Scholar on 31 July 2024.

cluster analysis, biomarker detection, metabolite annotation, and pathway analysis (functional analysis) on the processed data.

Figure 3(E–H) shows common downstream analyses that lead to the final goal of functional analysis. Overall, functional analysis comprises two primary modules: (i) metabolite or omics-level analysis (Fig. 3E–F), and (ii) pathway-level analysis and visualization (Fig. 3G–H). The details of each step in Fig. 3 are provided in Supplementary Note. The next section further details the downstream analysis pipeline.

Annotation and omics-level analysis

Figure 3(E–F) shows the common steps in annotation and omics-level analysis. The goal of this pre-analysis phase is to make the input data become adaptable for the functional analysis. The input can be either metabolite-by-sample matrix (targeted metabolomics) or feature-by-sample matrix (untargeted metabolomics). If the software does not have pathway information embedded, users need to provide the pathways in the .GMT format.

Metabolite annotation

Metabolite annotation is necessary for untargeted metabolomics data. Eight of the surveyed tools (XCMS, MetaboLyzer, Mummichog, MetDNA, MetaboAnalyst, specmine, metaX, and IP4M) support metabolite annotation using untargeted metabolomics data. Other tools require users to conduct putative metabolite identification for the input spectral features beforehand—a challenging task in the field [36]. A straightforward annotation approach involves searching the m/z values and the RTs of input features against a spectral library to identify potential chemical matches [61–63]. Notably, the effectiveness of this approach is limited by the completeness of the reference databases. To make the annotation more comprehensive, more sophisticated techniques have been developed to integrate multiple databases [64].

XCMS, specmine, metaX, and IP4M utilize third-party tools for metabolite annotation. XCMS uses METLIN [65]; metaX and IP4M use CAMERA [66]); specmine uses MAIT [67]. MetaboLyzer, MetaboAnalyst, Mummichog, and MetDNA implement their own search algorithms that basically match each input feature with known metabolites in spectral libraries (e.g. KEGG [17], HMDB [18], Lipid Maps [68], BioCyc [69]). MetaboLyzer, MetaboAnalyst, METLIN, CAMERA, and MAIT aim to find a metabolite that best matches the each feature whereas Mummichog searches for all tentative metabolites and then retains the metabolites that are locally close in the metabolic pathways. MetDNA retains the best-matched metabolite for each feature by utilizing tandem MS (MS2) spectral databases, assuming that two metabolites within a reaction pair will share similar MS2 spectra.

ID mapping

ID mapping involves converting omics IDs from input to the IDs used in a pathway database, e.g. KEGG Entrez ID or Compound IDs. Pathway databases typically describe a metabolic pathway

Method	Preprocessing & QC	Global Metabolomics	Input					Pathway Database	Metabolite Annotation	Cluster Analysis	Differential Analysis	Functional Analysis		Network Construction	Meta-Analysis
			G	T	E	P	M					F2P	M2P		
Over-Representation Analysis (ORA)															
XCMS	✓	✓					✓	KEGG, UCSD Recon1, EHMN	✓		✓	✓	✓		✓
PaintOmics				✓	✓	✓	✓	KEGG, Reactome, MapMan			✓		✓		
MBRole							✓	KEGG, HMDB, YMDB, PubChem, BioCyc, MeSH, ChEBI, ECMDB, TTD, CAS, PharmGKB, PathBank, Reactome, LIPID MAPs, ChemSpider					✓		
MetaboLyzer		✓					✓	KEGG, BioCyc	✓		✓		✓		
Mummichog		✓					✓	KEGG, UCSD Recon1, EHMN	✓		✓	✓			
MetDNA		✓					✓	KEGG	✓		✓		✓		
Functional Class Scoring (FCS)															
MetaboAnalyst	✓	✓	✓				✓	KEGG, UCSD Recon1, EHMN	✓	✓	✓	✓	✓	✓	✓
PAPi							✓	KEGG					✓		
IMPaLA				✓		✓	✓	KEGG, Reactome, BioCyc, PID, BioCarta, NetPath, INOH, EHMN, PharmGKB, WikiPathways, SMPDB					✓		
Pathview				✓			✓	KEGG					✓		
specmine	✓	✓					✓	KEGG, HMDB	✓	✓	✓		✓		
metaX	✓	✓		✓		✓	✓	KEGG, Reactome, BioCyc, PID, BioCarta, NetPath, INOH, EHMN, PharmGKB, WikiPathways, SMPDB	✓		✓		✓	✓	
Lilikoi							✓	KEGG, HMDB					✓		
MB-PLS							✓	KEGG					✓		
PUMA							✓	KEGG			✓		✓		
IP4M	✓	✓					✓	KEGG	✓		✓		✓		
M2IA							✓	KEGG			✓		✓	✓	
ogPLS							✓	KEGG					✓		
ssPA							✓	KEGG, Reactome					✓		
Topology-based Pathway Analysis (TPA)															
Metscape				✓			✓	KEGG, HMDB, EHMN			✓		✓	✓	
Omicsnet				✓		✓	✓	GO, GO-Slim, Reactome, KEGG			✓		✓	✓	
FELLA							✓	KEGG			✓		✓		
iMSEA							✓	KEGG					✓		
dci-MSEA							✓	KEGG					✓		

Figure 2. Capabilities of the reviewed software using metabolomics data. The first column (**Method**) lists the method names, while the second column (**Preprocessing & QC**) describes whether the methods can perform data processing and quality control using raw metabolomics data. The third column (**Global Metabolomics**) describes whether the methods support the analysis of untargeted metabolomics data (also referred to as shotgun/untargeted/nontargeted metabolomics). The next five columns under **Input** describe the supported input multi-omics types (**G**: Genomics, **T**: Transcriptomics, **E**: Epigenomics, **P**: Proteomics, and **M**: Metabolomics). Note that all methods reviewed in this article support the analysis of Metabolomics (i.e. the Metabolomics column is checked for all methods). The column **Pathway Database** lists the databases used by each method. The column **Metabolite Annotation** indicates whether the methods support metabolite annotation. The column **Cluster Analysis** indicates whether the methods support cluster analysis of the input samples. The column **Differential Analysis** indicates the methods support differential analysis and marker detection. The next two columns under **Functional Analysis** describes whether the methods are capable of performing functional analysis using the spectral features directly (**F2P**) or they need to transform the features into metabolites first (**M2P**). The column **Network Construction** indicates whether the methods also output a network constructed from the input data. The column **Meta-analysis** indicates the capability to perform meta-analysis (multi-cohort analysis).

using two main entities: compounds and their reactions. The reactions are catalyzed by enzyme(s) and/or protein(s). Therefore, metabolic pathways can be considered either as metabolite-centric networks focusing solely on their constituent metabolites, or as multi-omics networks consisting of metabolites interconnected with other omics entities. The former requires mapping input metabolites to the database's compound IDs, while the latter involves mapping genes, metabolites, and enzymes to their respective IDs. Pathway analysis methods either rely on user-provided mapping files or perform mapping using embedded databases, or a combination of both. Details regarding supported pathway databases and the ID mapping procedures for each method are provided in Supplementary Note.

Pathway augmentation

Pathway augmentation is the process of adding interactions among omics entities that are not fully represented in current pathway databases. For instance, a single gene can code for several distinct proteins due to the alternative RNA splicing [70]. Researchers also aim to incorporate the role of miRNAs and their targets into the analysis, but the ID mapping process often results in a one-to-many scenario, where an miRNA could target multiple genes or vice versa. Therefore, it is better to augment pathways to include all omics entities and their interactions instead of merging them into a single node. Researchers can enhance the comprehensiveness of current metabolic pathways by integrating validated or predicted multi-omics interactions from existing knowledge bases. PaintOmics [38, 39] and OmicsNet [56] employ this process in their analysis pipeline.

Pathway-level analysis

Figure 3(G) illustrates the common steps in pathway-level analysis, while Fig. 3(H) shows the outputs users can expect from

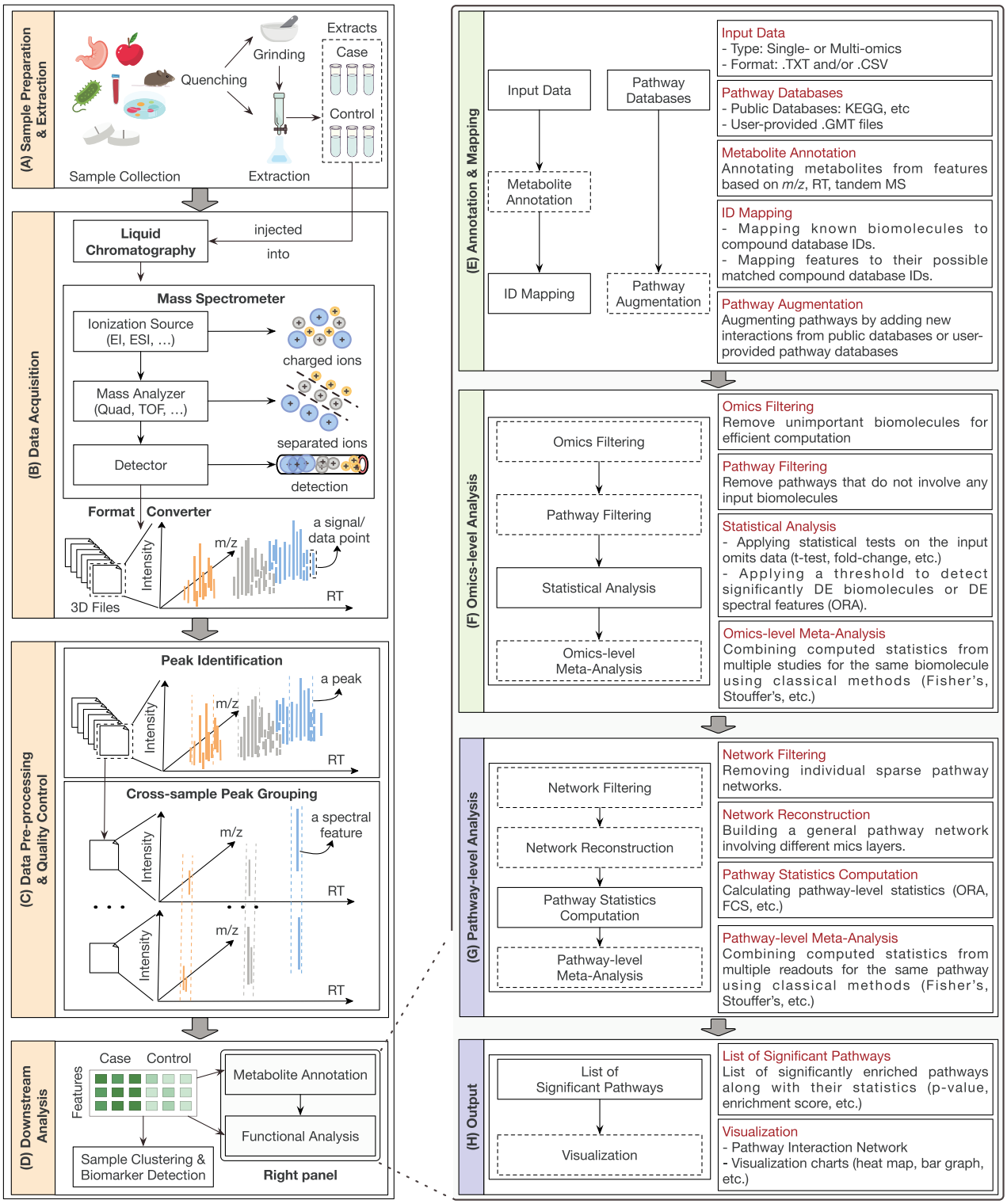

Figure 3. The complete workflow of metabolomics studies. The left panel describes four main steps in a metabolomics analysis pipeline using high-resolution LC-MS data: (A) sample preparation and extraction, (B) data acquisition using LC-MS instruments, (C) data processing and quality control, and (D) downstream analysis. The right panel further details the overall pipeline for downstream analysis: (E) metabolite annotation and mapping, (F) omics-level analysis, (G) pathway-level analysis, and (H) output and visualization. Solid-line boxes depict common modules in analysis pipelines, while dashed boxes represent optional modules.

the surveyed methods. The primary objective of pathway-level analysis is to calculate pathway-level statistics, such as enrichment scores, from omics-level data, and test for significant pathway enrichment. The output typically includes a table of significant pathways with their *P*-values, along with publication-ready visualizations, such as volcano plots or bar plots. In the following sections, we will discuss these steps for each method in detail. Specifically, we classify the methods into three categories: (i) ORA, (ii) FCS, and (iii) TPA. We then highlight notable characteristics of each category and present key points of each method.

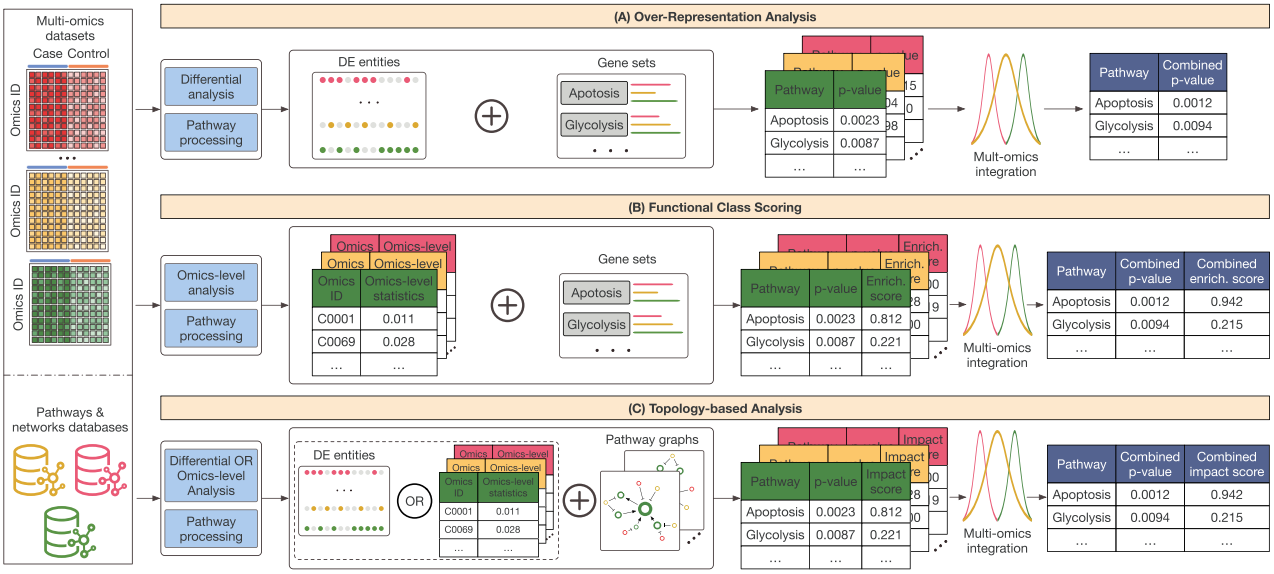

Figure 4. Three strategies for functional analysis: ORA (shown in panel A), (ii) FCS (panel B), and (iii) TPA (panel C). The input data of each method includes metabolomics data and pathway information (from KEGG, HMDB, BioCyc, etc.). Some methods allow users to integrate metabolites with other entities (e.g. genes, proteins, enzymes). (A) ORA: approaches in this category first identify differentially expressed (DE) entities and then uses hypergeometric test to identify pathways that have the number of DE entities more than expected. (B) FCS: approaches in this category first compute entity-level statistics (e.g. enrichment scores, t-statistic) and then aggregated them into pathway-level statistics for hypothesis testing to identify the enriched pathways. (C) TPA: approaches in this category combine ORA and FCS with network analysis techniques to identify impacted pathways.

Over-Representation Analysis

Figure 4(A) shows the overall workflow of ORA approaches. The underlying concept is that if the ratio of DE metabolites involved in a certain pathway surpasses the proportion expected by chance, then the pathway is over-represented [71, 72]. In multi-omics setting, ORA methods start with differential analysis on each readout to obtain omics-level statistics (*P*-values, log fold-changes, etc.) and then select the DE entities based on user-provided thresholds. Subsequently, ORA methods use hypergeometric test to compute the *P*-value of each pathway—one *P*-value per pathway per readout. To integrate results from multiple omics layers, ORA methods combine the omics-specific *P*-values for each pathway using a meta-analysis method (e.g. Stouffer's [73] or the weighted Fisher's method [74]). Finally, most methods adjust the *P*-values using false discovery rate (FDR) [75].

There are six ORA methods: XCMS [37], PaintOmics [38, 39], MBRole [40], MetaboLyzer [41], Mummichog [36], and MetDNA [42]. Among these, only PaintOmics can integrate multiple types of omics data: transcriptomics, metabolomics data, region-based omics (e.g. ChIP-Seq, DNase-Seq, ATAC-Seq, Methyl-Seq), and regulatory-based omics (e.g. miRNAs, transcription factors). PaintOmics first performs differential analysis to identify DE entities (metabolites, genes, peaks from ChIP-seq, etc.), followed by a hypergeometic test exact for each omics type. Subsequently, the PaintOmics adjusts the *P*-values before combining the *P*-values from omics layers to obtain a final list of *P*-values for the pathways.

Another method, MetDNA, requires as input a MS feature table (in.CSV format) and tandem MS (MS2) data files (in.mgf format). The tool also offers R functions to retrieve metabolic pathways and reaction pairs from KEGG. From input data, MetDNA identifies known metabolites and perform differential analysis (using *t*-test or Mann–Whitney–Wilcoxon test) before performing hypergeometric test to compute the *P*-values for the pathways.

Two methods, MBRole and MetaboLyzer, take as input a list of significant metabolites. MBRole embeds pathway information from 15 databases while MetaboLyzer embeds two databases (Fig. 2). The two methods perform hypergeometric test to compute the *P*-values for the pathways, and then adjust the *P*-values for multiple comparisons. MBRole also outputs the enrichment ratio for each pathway while MetaboLyzer also provides the list pathways with the highest occurrence of putatively identified metabolites and presents this information in a histogram for each database.

The last two approaches, Mummichog and XCMS, use the same algorithm for functional analysis (XCMS incorporates Mummichog's algorithm). Given the list of features, Mummichog matches each feature with all possible metabolites (background set) and then performs differential analysis to identify DE features. The selected features are also mapped into their possible matched metabolites. Next, Mummichog uses hypergeometric test to compute the *P*-values for the pathways. To adjust the *P*-values, Mummichog applies a permutation-based strategy [76] in conjunction with EASE score [77]. The EASE score for each pathway is obtained by subtracting one DE metabolite/feature from each pathway and then recalculating the *P*-value. The significance of the EASE score is estimated using the cumulative density function from the permutation, resulting in the adjusted *P*-value for each pathway.

The primary advantage of ORA approaches lies in the rapid prediction of major biological functions among massive datasets. However, these approaches bear the following three restrictions: (i) loss of important information due to the user-chosen cut-off method, (ii) disregard of interactions between biological molecules and pathways, and (iii) the assumption of independence among pathways. Exclusively, Mummichog with its permutation strategy addresses the third restriction.

Overall, ORA approaches are simple in term of methodology, which compare the list of DE features against the background features using hypergeometric test (or Fisher's exact test) to determine the pathways that are significantly enriched [78]. Therefore, users can rapidly obtain the prediction of major biological

functions among massive datasets [71]. However, these approaches bear the following three restrictions: (i) loss of important information due to the user-chosen cut-off method [79], (ii) disregard of interactions between biological molecules and pathways [80], and (iii) the assumption of independence among pathways. Exclusively, Mummichog with its permutation strategy addresses the third restriction [81].

Functional Class Scoring

Figure 4(B) shows the general workflow of FCS approaches. The main difference between FCS and ORA is that FCS eliminates the step of identifying DE metabolites or other omics entities. It means that all input compounds can be incorporated in the functional analysis. The general FCS workflow consists of the following steps: (i) computing omics-level statistics, (ii) aggregating omics-level statistics to pathway-level statistics, (iii) hypothesis testing to identify enriched pathways, and (iv) combining *P*-values and enrichment scores from multiple omics layers for each pathway. Most methods adjust the *P*-values for multiple comparisons using FDR [75].

There are 13 FCS methods: MetaboAnalyst [43], PAPi [28], IMPaLA [30], Pathview [44], specmine [45], metaX [46], Lilikoi [47], MB-PLS [48], PUMA [49], IP4M [50], M2IA [51], ogPLS [52], and ssPA [53]. Following the data processing of XCMS, MetaboAnalyst integrates various statistical analyses, covering metabolite-level analyses, functional analysis, network construction, and meta-analysis at both biomarker and functional levels. MetaboAnalyst initially incorporated Mummichog's ORA algorithm, but the authors later on extended the Mummichog algorithm to an FCS algorithm implemented in GSEA [82].

PAPi and M2IA use the same six-step pipeline to generate a pathway score matrix. PAPi first ranks pathways based on the total number of their constituent metabolites. For each patient, the method calculates the sum of the measurements of the constituent metabolites for each pathway. Next, it divides the total measurement by the ratio of pathway rank to its size, resulting in a pathway activity score matrix (pathway by patient). Finally, PAPi performs *t*-test and ANOVA to compute the *P*-values of the pathways. M2IA embed PAPi's algorithm for functional analysis, but it also integrates other tools for microbiome analysis, including data processing, statistical analysis, network analysis, and functional analysis.

Lilikoi uses a third-party tool named Pathifier [83] to construct a Pathway Deregulation Score matrix of pathways for each input sample. Next, the method filters out pathways not relevant in separating sample groups (disease versus control). Finally, Lilikoi provides users with seven machine learning models capable of classifying the samples their respective groups (disease versus control). Lilikoi estimates the importance scores of the pathways according to their contribution to the sample classification.

MB-PLS and ogPLS use a strategy similar to that of Lilikoi. MB-PLS creates a block for each KEGG pathway which is a matrix where rows represent samples and columns represent metabolites belonging to the pathway. Next, MB-PLS concatenates the columns of all matrices and then builds a partial least squares model capable of classifying the samples to their respective groups. Based on the classifier, MB-PLS estimates the importance score of each metabolite and then calculates the importance score of each pathway as the weighted sum of its constituent metabolites. The other method, ogPLS, enhances MB-PLS by combining it with a lasso technique that penalizes metabolite participating in multiple pathways [84, 85]. The method also performs resampling to assess how each pathway contributes to the stability of the model [86]. The method outputs significant pathways and their importance scores.

ssPA also uses block matrices but it utilizes two alternative methods for computing pathway score matrix: k-means clustering (ssClustPA) and kernel PCA (kPCA). Given a block, ssClustPA clusters the block using k-means with $k = 2$ and then calculates the column vector that represents the difference of two centers. The method then multiplies the block with the difference vector to obtain a column vector that represents the score of the pathway across the samples. The same is procedure is applied for all blocks to compute the pathway score matrix. kPCA is similar to ssClustPA, except it applies kernel PCA [87] with a radial basis function kernel to compute the scores. Finally, ssPA performs a *t*-test on the pathway score matrix to compute the *P*-value for each pathway.

IMPaLA supports the integration and functional analysis using transcriptomics, targeted proteomics, and targeted metabolomics. IMPaLA allows users to choose between ORA and Wilcoxon Enrichment Analysis (WEA) [88]. For WEA approach, IMPaLA first calculates the summary statistics at the omics layers and then uses Wilcoxon signed-rank test to calculate the *P*-values for each pathway for each omics data type. Finally, the *P*-value of each pathway is calculated as the product of the *P*-values obtained from all omics types. The tool outputs the combined *P*-values and the *P*-values obtained for each data type.

PUMA constructs a Bayesian network that takes as input pathways and metabolites with their m/z value as prior knowledge. PUMA then defines a conditional probability of the status of pathways given their prior probability of being active (following either a Bernoulli or a Beta distribution). Finally, Gibbs sampling is used to perform Bayesian inference, approximating the posterior probability of each pathway being active, given the m/z values of its constituent metabolites.

Pathview uses different techniques for the functional analysis depending on the input data. When users provide a list of DE genes or metabolites, Pathview uses ORA. When users provide the full matrices, Pathview uses a third-party tool named GAGE [89]. The methodology of GAGE involves conducting metabolite set testing to assess the enrichment of metabolites in a particular pathway. The significance of the enrichment is then evaluated using a permutation-based approach. Pathview visualizes resulting pathways as the native KEGG pathway and Graphviz [90] view. The Graphviz view offers enhanced control over node/edge attributes and provides a more detailed representation of graph topology.

IP4M, specmine, and metaX embed many third-party tools in their software. IP4M includes 62 different functions for data processing (using XCMS and CAMERA), basic statistical analysis, classification and biomarker detection, correlation analysis, cluster and sub-cluster analysis, regression analysis, ROC analysis, functional analysis (using MetaboAnalyst), and power analysis. metaX also embeds many tools for the analysis of untargeted metabolomics data, including data processing (using XCMS and CAMERA), missing value imputation (using kNN, missForest [91], etc.), data normalization (using combat [92], quantiles, etc.), and functional analysis (using IMPaLA). Similarly, specmine utilizes third-party tools for metabolomics data analysis, including data processing (using XCMS), metabolite annotation (using MAIT), and regression (using models implemented in caret). To perform the functional analysis, specmine calculates the *P*-values and statistics for the input metabolites and then performs a GSEA-like analysis to calculate the *P*-values of the pathways.

Overall, FCS has several advantages over ORA. It performs functional analysis using the measurements of all input metabolites and entities (instead of focusing on differential expressed ones). This allows users to analyze their data without imposing an arbitrary cutoff to select DE entities [80, 93]. In turns, it increases the reliability and reproducibility. Another advantage is that FCS is capable of taking into account the dependency between the metabolites and the coordinated changes of all entities. However, FCS still does not take into consideration pathway topology and pathway crosstalk [94]. FCS methods often assign equal weight to all pathways and neglect the relationship between pathway components and other aspects of the network structure of pathways [80, 95]. Moreover, FCS methods analyze pathways independently, ignoring the overlaps between them and the potential influence one pathway can exert on another [81].

Topology-based Pathway Analysis

Figure 4(C) shows the general TPA workflow. The main difference between TPA from FCS and ORA is that TPA approaches take into consideration both the expression change of the multi-omics data and pathway topology. TPA methods basically performs network analysis using the summary statistics obtained from ORA or FCS approaches. Here we review five TPA approaches: Metscape [54, 55], Omicsnet [56], FELLA [57], iMSEA [58], and dci-MSEA [59].

MetScape uses LRpath [96] for functional analysis, which basically performs logistic regression to identify the pathways that are significantly enriched. Through Cytoscape [97], MetScape can visualize different types of networks from HUMDB [98], KEGG [17], and EHMN [99]: Compound-Reaction-Enzyme-Gene (CREG) network graph, Gene-Compound (GC) network graph, Compound-Reaction (CR) network graph, and Compound (C) network graph. If users perform GSEA [82] on their own, they can provide the GSEA results to MetScape for visualization.

OmicsNet supports users to construct networks for three types of interactions: protein–protein, microRNA–gene, transcription factor–gene, or protein–metabolite, with each representing a separate layer. For functional analysis, the tool uses hypergeometric test on genes/metabolites within the generated interaction network. This analysis is conducted against GO, PANTHER GO-Slim, Reactome, or KEGG pathways, resulting in a list of pathways or GO terms that are impacted significantly. OmicsNet also facilitates three analyses of network topology analysis: node centrality analysis, module detection, and identification of the shortest paths.

FELLA takes as input a list of DE metabolites and performs functional analysis using ORA or TPA. For network analysis, FELLA scores the nodes within individual pathway networks using two alternative algorithms: PageRank [100] and heat diffusion model [101]. PageRank utilizes the random walks algorithm to score nodes and pathways, while the heat diffusion model performs sub-network analysis on the networks to extract a meaningful subgraph. Finally, FELLA performs permutation (using Monte Carlo trials or z-test) to compute the *P*-values of the pathways.

iMSEA focuses on identifying impacted pathways for patients taking different drug treatments. iMSEA first applies an algorithm named Partial Least-Squares Discriminant Analysis (PLS-DA) [102] to select metabolites that contribute to treatment differences. Next, iMSEA performs a random walk with restart to score the nodes and calculates pathway activity scores. To determine the significance of altered pathways, iMSEA performs a permutation test using the activity scores. The method also calculates a combination index for each pathway that indicates the interaction of different drugs (e.g. synergy, additivity, or antagonism).

dci-MSEA is considered an enhanced version of iMSEA. dci-MSEA first reconstructs a general metabolite network where each node represents a metabolite and each edge is a reaction pair. Next, dci-MSEA constructs a correlation network of DE metabolites in which edge weights are correlation coefficients and node weights are intensity levels. Next, it integrates the general metabolite network with correlation network to construct a differential correlation-informed metabolite network (dci-metabolite network). Next, it performs a random walk-based propagation on the dci-metabolite network to score metabolites iteratively, followed by a pathway-level activity score calculation. Finally, *P*-value dci-MSEA calculates pathway *P*-values one-sample z-test with permutation testing.

TPA is the latest generation of pathway approaches in the field, aiming to address all restrictions from the two previous generations. TPA analysis leverages topology information to account for biologically relevant differences between components, assigning more weight to changes in genes with greater influence over the pathway [78, 103]. By incorporating topological information, TPA analysis allows for a more precise examination of the same set of pathway components, recognizing that interactions may vary under different biological conditions [104, 105]. Additionally, TPA analysis considers causal interactions within pathways, acknowledging that modifications in upstream components can alter the behavior of downstream components [71]. However, TPA has its own limitations, such as insufficient consideration of interactions among pathways or the difficulty in constructing an unbiased network for network topology analysis [106]. A recent holistic assessment by Lu *et al*. [35] suggests that TPA approaches are not always be the best choice for functional analysis. The study recommends a combination of NetID [107] (an LC-MS peak annotation tool) and Mummichog instead.

Summary and practical guideline

Here we provide critical interpretation and analysis of all methods to help readers in their research purpose. We also implement wrappers of all standalone tools and make them available at https://github.com/tinnlab/metabolite-pathway-review-docker. This will allow users to download and execute the software with ease.

First, we summarize all relevant information discussed thus far in Supplementary Table S1. The table summarizes important information about the reviewed methods: Preprocessing & QC, Global Metabolomics, Input, Pathway Database, Metabolite Annotation, Cluster Analysis, Differential Analysis, Functional Analysis, Network Construction, Meta-Analysis. The column Key Technique indicates the core technique for functional analysis. The two columns Strengths and Weaknesses detail the advantages and disadvantages of each individual method, respectively. The column Usage provides a detailed description on how users can execute each method. For web-based tools, the column points to the tutorial page of each tool. For standalone tools, the column describes how users can run the methods using our newly created wrappers.

Second, we provide an evaluation of each method in Supplementary Table S2. We use the following metrics: software accessibility, installation, documentation, and user-friendliness. Each metric is scored on a scale from 1 to 10, with higher scores indicating better evaluation. Regarding software accessibility, most web-based tools are scored high due to their ease of access and the seamless nature of updates, which do not necessitate software reinstallation by users. The installation metric measures how easy to install the software. MetaboLyzer, PAPi, metaX,

and ssPA require manual installation of dependencies without guidance while ogPLS and PUMA need source code fixes. The documentation metric assesses the quality of documentation, including working examples, parameter setting, and installation instructions. User-friendliness rates software design from user perspectives. Web-based and GUI tools that are easy to navigate receive the highest scores.

Finally, we provide a general guideline (Fig. 5) for selecting the most suitable functional analysis methods in specific situations. To demonstrate how to use the guideline to select the best method for a specific application, consider a scenario where users need to preprocess 3D files. In this case users can choose any of the five methods: XCMS, MetaboAnalyst, specimine, IP4M, and metaX. Among these, MetaboAnalyst has the highest score (a total score of 40) while metaX allows users to integrate metabolomics data with other omics types. Note that besides experiment design, input, and evaluation scores, users can also choose a method based on methodological categories (ORA, FCS, TPA).

Outstanding challenges and solutions

This section discusses the important challenges in the field that remain unsolved. Here we compose a comprehensive list of these challenges in Fig. 6(A–D), emphasizing future functional analysis tools should take these factors into consideration. These include challenges in LC-MS data processing, metabolite annotation, incomplete pathway databases, and methodological limitations. Although LC-MS data processing and metabolite annotation are general challenges for metabolomics studies, they have significant impacts on functional analysis. Proper data processing helps to reduce noise, correct for instrumental variations, and improve the overall signal-to-noise ratio. This is crucial for detecting true biological signals that directly influence functional analysis results [35, 108]. Similarly, metabolite annotation is crucial for functional analysis because most approaches take the list of metabolites and their intensity as input. Accurate annotation allows researchers to map detected metabolites to known biological pathways and functions, enabling a deeper understanding of the underlying biological processes [109].

Parameter setting for LC-MS data processing

The quality of the processed data affects the accuracy of subsequent analyses, including metabolite annotation, biomarker detection, and functional analysis [35]. In each run, LC-MS generates a 3D file that contains critical information such as retention time (RT) data in chromatograms, mass-to-charge ratio (m/z) in MS spectra, and the relative abundance for each specific ion. To extract meaning insights from these 3D files, a robust preprocessing pipeline is essential to generate meaningful features (i.e. peaks) characterized by their RTs, m/z values and intensities at a low false-positive rate. Pivotal stages in this pipeline encompass noise filtering, peak detection, RT alignment, peak matching, and normalization.

Suboptimal parameter choices can easily lead to biased results, thereby impacting subsequent analyses. Various software solutions have emerged to facilitate users in executing this pipeline, spanning both commercial and non-commercial domains. Among the tools that can perform functional analysis, XCMS stands out as the most popular choice, with a majority of the methods relying on its capabilities for data processing. There are other tools that were recently developed specially for optimizing the LC-MS Data Processing, including IPO [110], AutoTuner [111], MetaboAnalystR [112], DeepRTAlign [113], and PeakDecoder [114]. IPO, AutoTuner, MetaboAnalystR were designed to optimize the parameters in the process used for XCMS. DeepRTAlign is a deep learning-based tool trained on multiple datasets to improve RT alignment in LC-MS studies. PeakDecoder learns patterns from raw data to accurately identify when different compounds are genuinely eluting or moving together through the chromatographic and MS processes.

Despite recent efforts, choosing optimal parameters remains a challenge due to the vast number of parameter combinations and the difficulty in measuring the quality of the results. One direction to improve the quality of LC-MS processing is to introduce more experimentally acquired chromatography retention time datasets, including both monotonic and non-monotonic RT shifts. This would allow us to leverage the power of deep learning models to learn patterns in the data more effectively and to have more resources to validate the capabilities of these methods. Additionally, although DeepRTAlign primarily focuses on MS data for RT alignment, it does not utilize tandem MS (MS2) information. MS2 involves the breakdown of selected ions (precursor ions) into ion fragments (product ions). Incorporating this MS2 information, such as ion intensities and fragmentation patterns, could enhance the model's accuracy. A potential approach is to extend DeepRTAlign into a multi-task learning framework. In this framework, one task would handle MS data-based RT alignment, while the other would focus on MS/MS (MS2) data.

Incomplete metabolite annotation

Identifying the set of putative metabolites (metabolite annotation) plays a crucial role in functional analysis because they are the direct input of most functional analysis methods. High-resolution LC-MS instruments in untargeted mode can yield more than 10,000 peaks for human samples. Only a fraction of these peaks can be confidently annotated, leaving a substantial percentage of peak unidentified [115]. Many peaks may represent in-source fragments, adducts, and isotopes of the same metabolite, thereby complicating the identification of the complex metabolite pool.

Another important matter is the accuracy of the annotation. Mass tolerance of the LC-MS instruments plays an important role in this matter [35, 116, 117]. Traditionally, many believe that a mass tolerance of 5 ppm or less would suffice [118], but this remains controversial. However, a recent benchmark article suggests that the ideal mass tolerance should be between 1 and 3 ppm, especially for annotation approaches that are based on m/z values or a combination of m/z values and RTs [35]. Other articles indicate that a very high mass accuracy (e.g. less than 1 ppm or even 0.1 ppm) is still insufficient for observing accurate outcomes of metabolite annotation [116, 117].

Many methods have been developed to facilitate a comprehensive annotation of metabolites. These include deep learning approaches that train their models on millions of known chemical structures to predict compound classes given spectral data [119–121]. Other methods employ network analysis for metabolite annotation, wherein known metabolites within a cluster of connected peaks are used in annotating their neighboring compounds [107, 122, 123]. Despite recent efforts, the predictive abilities and accuracy remain significantly constrained by our current knowledge bases [124–126].

Recognizing the limitations of current annotation approaches, Mummichog attempts to reduce the need for a complete annotation. Mummichog assigns all matched metabolites to DE features, and keeps pathways that have those metabolites working locally in concert. One drawback is that the assumption on locally connected subgraphs might not hold true in all cases,

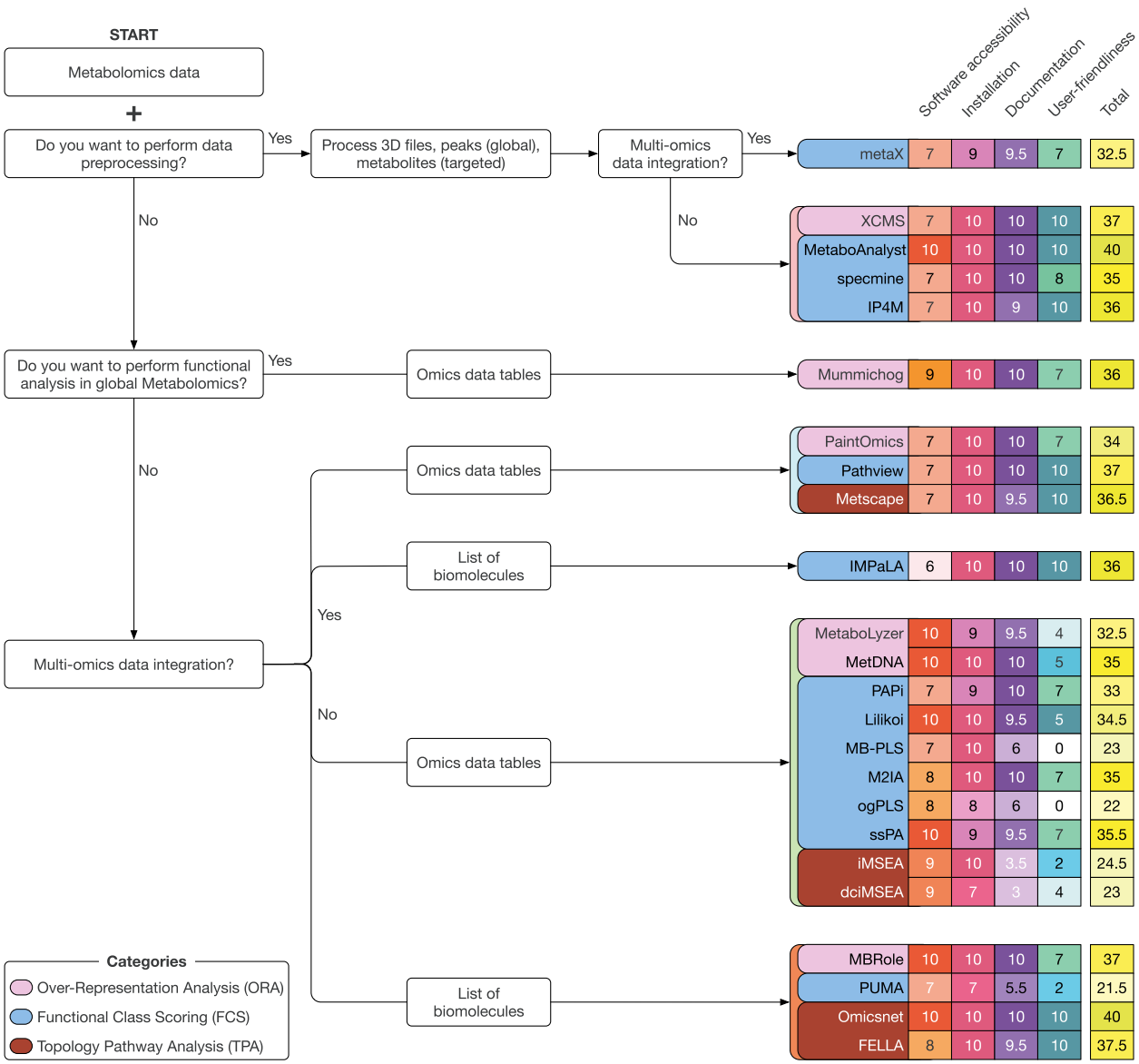

Figure 5. Guideline for selecting suitable methods for functional analysis. Depending on research design and input, users can follow the arrows on the diagram to choose a suitable approach. For instance, users who wish to preprocess metabolomics data (e.g. 3D files), then they can choose among five methods: XCMS, MetaboAnalyst, specmine, IP4M, and metaX. XCMS is ORA-based while the remaining four methods are FCS-based. If users wish to preprocess and integrate metabolomics with other omics types then metaX is the only option. If users have global/untargeted metabolomics data, they can use Mummichog to perform functional analysis. Otherwise, they can choose from the remaining 18 methods. Among these, four methods (PaintOmics, Pathview, Metscape, and IMPaLA) can perform functional analysis using multi-omics integration, while the rest only work with metabolomics data. The diagram also shows the summary statistics of the methods, including the quality of their ease of access, ease of installation, documentation, and user-friendliness. Users can also choose a method based on the statistics provided, or based on method category (ORA, FCS, TPA).

i.e. pathways might be active even when constituent metabolites are not strongly connected [127]. This algorithm is also used by XCMS and MetaboAnalyst. MetaboAnalyst further improves Mummichog by combining both known metabolites and undetected peaks in their functional analysis. MetaboAnalyst first performs annotation to tentatively identify a set of metabolites with high confidence, leaving a subset of undetected peaks. It then utilizes Mummichog's algorithm to perform functional analysis using both sets.

There are two strategies to potentially mitigate the impact of incomplete metabolite annotation on functional analysis. One strategy is to combine the knowledge available in different annotation databases. Such approach would allow us to increase the number of identified metabolites and the reliability. Another strategy is to further improve the approach introduced by MetaboAnalyst, i.e. combining both identified and undetected peaks in the process of pathway analysis and functional annotation. We can combine multiple databases and methods to identify a reliable set of known metabolites and a set of peaks that cannot be detected by any method or database. After that, we can use different methods for functional analysis using both detected metabolites and undetected peaks.

Incomplete pathway databases

Notably, there is a limited coverage of experimentally validated metabolites in the current pathway databases, such that roughly half of known compounds can be found in pathway databases [128]. The metabolic functions of missing compounds remain

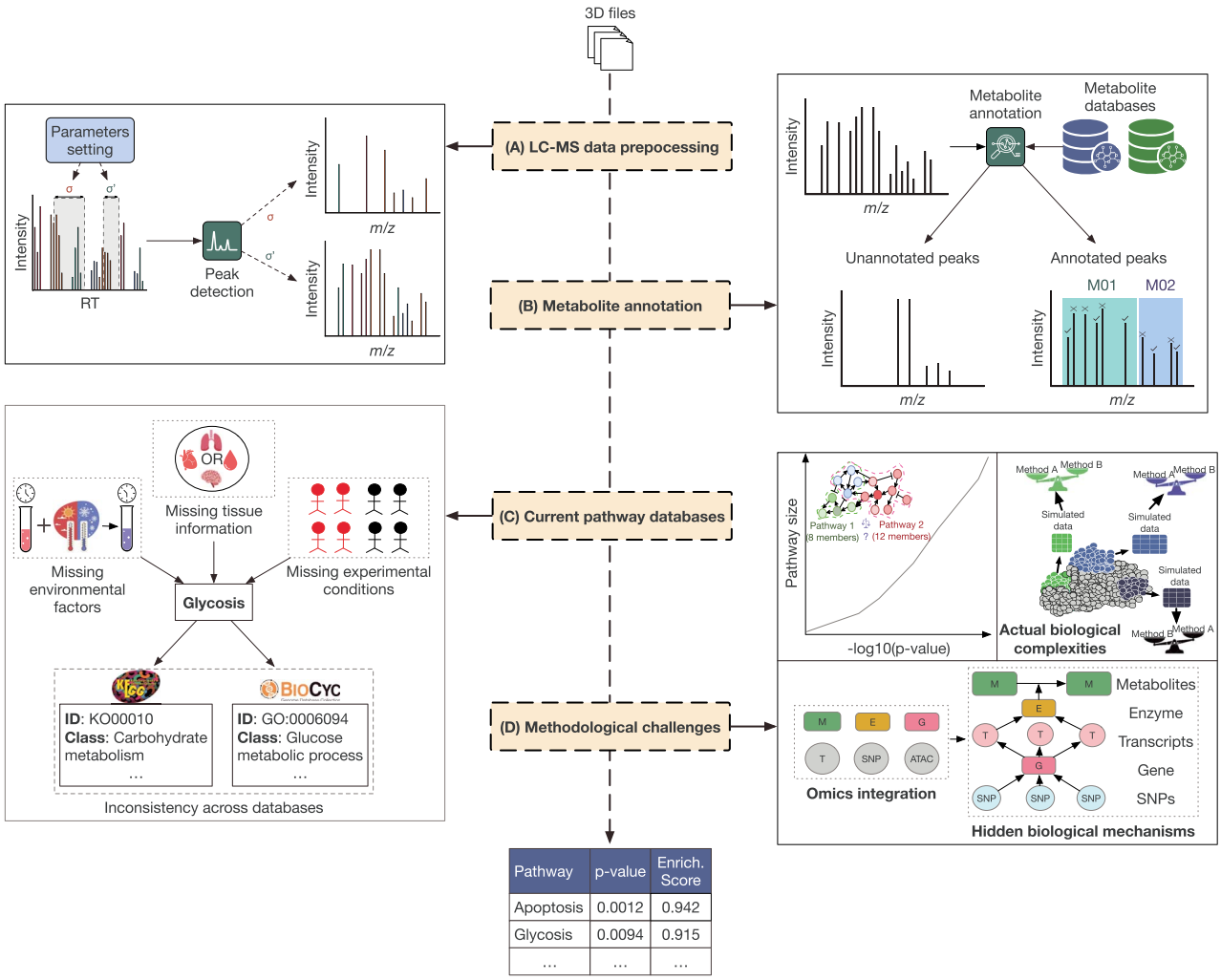

Figure 6. Outstanding challenges in functional analysis using metabolomics data. The challenge arises at each stage of the analysis pipeline, spanning from the initial processing of 3D files from LC-MS instruments to the identification of significantly impacted pathways. (A) LC-MS data preprocessing requires optimal parameter selection to effectively extract meaningful spectral features. (B) Metabolite annotation poses a substantial computational challenge due to the scarcity of comprehensive reference metabolite databases. (C) The discrepancies and gaps in current pathway databases affect the performance of functional analysis methods. (D) Limitations in methodology include a bias towards well-established pathways, the need for an optimal approach to integrate multi-omics data and the lack of benchmarking datasets for method assessment.

elusive due to the incomplete nature of pathway databases, which lack comprehensive information on known biochemical reactions [17]. Therefore, enhancing pathway curation in which unassigned compounds are involved stands as important area of scientific investigation [129].

Another critical considerable effect is the type of compound identifiers utilized in the pathway databases. For example, KEGG and BioCyc use their own identifiers, while Reactome uses ChEBI identifiers. Converting data identifiers to database-specific equivalents may result in information loss and ambiguous mapping, as not all identifiers map directly to database IDs. Many computational solutions have been proposed to bridge the dataset elements and pathway databases nodes based on chemical ontologies [130, 131]. However, this may affect pathway analysis results, as multiple data elements may map to a single node in the pathway database. Moreover, a discrepancy often exists between the chemical descriptions in pathway databases and the datasets. A study by Mubeen *et al.* [132] demonstrates that an integrated resource could yield more reliable outcomes compared to using a single database in isolation. Hence, further investigation is required to explore the interplay between chemical ontologies across databases, enabling the development of a unified source for compound identifier mapping.

In response, data-driven approaches could be leveraged to identify novel pathways or to expand existing ones. Recent efforts include ML techniques, such as graph or network embedding techniques [133] and link prediction frameworks [134], that could be applied to metabolomics data to identify metabolite-metabolite associations and predict missing pathway interactions. Graph embedding techniques enable the representation of complex biological networks in lower-dimensional spaces while preserving the structural information, allowing for more efficient analysis and the identification of hidden metabolite relationships that might not be apparent in raw data. Link prediction frameworks, by leveraging paths of length three (L3), identify potential interactions between metabolites that are not directly connected but share common neighbors, making it particularly useful for filling gaps in existing networks and expanding known metabolic pathways.

To address the challenge of incomplete pathway databases, a key strategy could be to actively expand the collection of metabolic pathways from natural sources. This would involve concerted efforts to discover new pathways linked to

well-established metabolites, as well as to identify new metabolites, particularly those from under-explored ecosystems that may contribute to new pathways. Integrating these new pathways into existing databases would enhance make pathway databases more comprehensive. In addition, one can consider adding the interaction among multi-omics data to make the pathways more complete. By integrating different omics layers, investigators can identify novel pathway connections that are relevant to the studied conditions.

It is also worth mentioning that experimental conditions, cell types, and time points associated with pathway construction are often unspecified in pathway databases. Biochemical reactions can be influenced by environmental factors at various time points, resulting in diverse outcomes when identifying affected pathways. Therefore, one can consider adding these information to pathway databases and pathway analysis approaches. Integrating information on cell and tissue types, diverse experimental conditions, and environmental exposures into pathway interactions will be essential areas of investigation for future method development.

Methodological limitations

There are a number of measurable and unmeasurable factors that lead to bias in functional analysis. Inherent disparities between pathway databases inevitably influence the results of functional analyses, regardless of the method employed [135]. Notably, the size of pathway can introduce bias in statistical analysis (e.g. ORA and FCS). As such, larger pathways are more likely to be identified as significant by enrichment analysis methods [35]. Among the surveyed methods, ogPLS is the only approach that resolves this issue by introducing a pathway debiasing coefficient to penalize pathways sizes. Another issue is that the number of pathways involved in an analysis directly can influence the adjusted *P*-values, i.e. testing more pathways leads to a greater loss of statistical power [78].

Among the examined methods, TPA-based approaches offer partial correction for bias. However, accounting for the intricate interactions among biochemical reactions can significantly complicate pathway analysis methods. Metabolites interact dynamically, in which output from one reaction serving as input for another. Furthermore, pathway databases like KEGG, Reactome, or BioCyc encompass both metabolic and signaling pathways, each of which represents pathway components differently. Metabolic pathways represent chemical reactions in which the metabolites are substrates and products while enzymes are catalysts. These chemical reactions can be represented as a graph in which nodes are metabolites and edges are reactions that can be boosted by enzymes (located in the middle of the edges). In contrast, signaling pathways use nodes to represent genes/compounds and edges to represent the interactions among compounds.

Because of the distinctively different structures between metabolic and signaling pathways, TPA methods developed for signaling pathways may not be able to analyze metabolic pathways and vice versa. For instance, both CrossTalkZ [136] and SPATIAL [137] require to exclude metabolic pathways before constructing their gene interaction network. This limitation possibly stems from relying only on single-omics data. To mitigate this issue, software developers should focus on integrating genomics and metabolomics information to provide a more comprehensive view of pathway activity. In addition, it is important to note that current methods (ORA, FCS, TPA) often disregard metabolite–metabolite correlations within pathways, which might introduce bias to the pathway results [35, 138]. Integrating a preprocessing step that filters out one metabolite from each highly or lowly correlated pair might be effective. This approach would focus the functional analysis on metabolites with moderate and biologically relevant interactions.

Multi-omics data integration also poses significant challenges in method development [139]. First, different omics data require tailored and specific data processing procedures that require different levels of expertise. Second, certain omics layer might be important in a condition but is relevant in another. However, current integrative methods over simplify the integration process and often weight all omics layers the same in all conditions. Third, it is essential to consider inter-dependency among metabolites and other multi-omics layers. Incorporating such inter-connectivity in functional analysis would greatly enhance statistical power and interpretability of analysis results. For example, one potential improvement is to should model the impact of single-nucleotide polymorphisms (SNPs) on pathways, which is completely missing in current knowledge bases. In addition, one gene might code for multiple transcripts that lead to different proteins and enzymes, which should be considered in functional analysis. Recently, MetaboAnalyst proposed a causal analysis between genomics and metabolomics with phenotypes to detect potential metabolite-phenotype associations [140, 141]. This paves the way for integration of SNPs and metabolomics data in the future. However, the proposed method is oversimplified and there are many other omics layers that need to be accounted for.

Further, when it comes to method assessment, there is a notable absence of benchmark datasets where we know exactly which pathways are impacted and which are not. Due the lack of real benchmark datasets, functional analysis approaches are often evaluated using simulated data. The common practice in generating simulation includes the following: (1) add noise to real metabolomics data (targeted and untargeted), (2) knockout enzymes in specific pathways, or (3) alter the status of reactions [78]. Although this approach has the ability to simulate a large number of samples, it may not reflect real-world scenarios. In addition, simulation is subjected to bias because simulated data is generated based on some assumptions which are usually identical with the assumptions made in designing the approach. Consequently, opting for real biological data as benchmark datasets is preferable for a more nuanced evaluation.

In summary, several challenges remain despite the progress made in functional analytical methods. First, selection of parameters for peak preprocessing, metabolite annotation, and absence of context- and cell-specific information are out-of-methodological hurdles. These issues indirectly present barriers to the implementation of functional analysis methods at their full potential. Second, many methodological assumptions that are present during development of tools cause variations in the outcome of a functional analysis approach. To harness the potential benefits of high-throughput technologies and enhance our understanding of large biological systems, the community needs to collaboratively address these challenges to advance functional analysis, enhancing its specificity, sensitivity, and relevance. In conclusion, functional analysis approaches for metabolomics data have progressed significantly in the past two decades, providing deep insights into the mechanisms underlying biological-relevant phenotypes.

Conclusion

Over the last two decades, functional analytical methods have served as the primary means for uncovering mechanisms

underlying different phenotypes. This review summarizes the state of functional analysis approaches designed for metabolomics data. The field aims to identify pathways that are perturbed by a condition through computational approaches. We examined the three main categories of methods: ORA, FCS, and TPA. Each category has strengths and limitations, and the elected method depends on the research question and the available data. ORA methods are computationally efficient and provide a simple way to identify overrepresented pathways. However, these methods rely on metabolites differentially expressed and do not integrate the magnitude of expression changes. FCS methods address this limitation by using the quantitative data and changes of metabolites within the pathways. TPA methods incorporate topological information of metabolic networks, leading to biologically meaningful results. Despite the long-standing proposal of these categories, current approaches assigned to them have not satisfactorily addressed the existing weaknesses.

Key Points

- The article recapitulates the complete workflow of metabolomics experiments.
- Functional or pathway analysis holds significance as it offers understanding into the biological mechanisms beyond biomarker detection and differential analysis.
- This article thoroughly investigates 24 pathway analysis approaches in terms of their accessibility, supported databases, data types, methodologies, and other downstream analyses.
- The article discusses outstanding challenges in metabolomics studies that need to be addressed by future research.
- The main objective is to provide experimenters and potential users with a complete picture of available resources so that they can choose the most appropriate approach for their research goal.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This work was partially supported by NSF (grant no. 2343019 and 2203236), NCI (grant no. 1U01CA274573-01A1), and NIGMS (grant no. 5U54GM104944 and 1R44GM152152-01). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Aguilar-Ramirez D, Herrington WG, Alegre-Díaz J. *et al*. Adiposity and NMR-measured lipid and metabolic biomarkers among 30,000 Mexican adults. *Commun Med* 2022;**2**:143. https://doi.org/10.1038/s43856-022-00208-2.
2. Fizelova M, Miilunpohja M, Kangas AJ. *et al*. Associations of multiple lipoprotein and apolipoprotein measures with worsening of glycemia and incident type 2 diabetes in 6607 non-diabetic Finnish men. *Atherosclerosis* 2015;**240**:272–7. https://doi.org/10.1016/j.atherosclerosis.2015.03.034.
3. Ahola-Olli AV, Mustelin L, Kalimeri M. *et al*. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia* 2019;**62**:2298–309. https://doi.org/10.1007/s00125-019-05001-w.
4. Soininen P, Kangas AJ, Würtz P. *et al*. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* 2015;**8**:192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216.
5. Würtz P, Havulinna AS, Soininen P. *et al*. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* 2015;**131**:774–85. https://doi.org/10.1161/CIRCULATIONAHA.114.013116.
6. Martínez-Reyes I, Chandel NS. Cancer metabolism: Looking forward. *Nat Rev Cancer* 2021;**21**:669–80. https://doi.org/10.1038/s41568-021-00378-6.
7. Lin L, Tang Y, Ning K. *et al*. Investigating the causal associations between metabolic biomarkers and the risk of kidney cancer. *Commun Biol* 2024;**7**:398. https://doi.org/10.1038/s42003-024-06114-8.
8. Tynkkynen J, Chouraki V, van der Lee SJ. *et al*. Association of branched-chain amino acids and other circulating metabolites with risk of incident dementia and Alzheimer's disease: a prospective study in eight cohorts. *Alzheimers Dement* 2018;**14**:723–33. https://doi.org/10.1016/j.jalz.2018.01.003.
9. Shen B, Yi X, Sun Y. *et al*. Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 2020;**182**:59–72.e15. https://doi.org/10.1016/j.cell.2020.05.032.
10. Yen NTH, Anh NK, Jayanti RP. *et al*. Multimodal plasma metabolomics and lipidomics in elucidating metabolic perturbations in tuberculosis patients with concurrent type 2 diabetes. *Biochimie* 2023;**211**:153–63. https://doi.org/10.1016/j.biochi.2023.04.009.
11. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* 2016;**17**:451–9. https://doi.org/10.1038/nrm.2016.25.
12. Ahadi S, Zhou W, Rose SMS-F. *et al*. Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat Med* 2020;**26**:83–90. https://doi.org/10.1038/s41591-019-0719-5.
13. Buergel T, Steinfeldt J, Ruyoga G. *et al*. Metabolomic profiles predict individual multidisease outcomes. *Nat Med* 2022;**28**:2309–20. https://doi.org/10.1038/s41591-022-01980-3.
14. Want EJ, Cravatt BF, Siuzdak G. The expanding role of mass spectrometry in metabolite profiling and characterization. *Chembiochem* 2005;**6**:1941–51. https://doi.org/10.1002/cbic.200500151.
15. Saghatelian A, Trauger SA, Want EJ. *et al*. Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* 2004;**43**:14332–9. https://doi.org/10.1021/bi0480335.
16. Fiehn O, Kopka J, Dörmann P. *et al*. Metabolite profiling for plant functional genomics. *Nat Biotechnol* 2000;**18**:1157–61. https://doi.org/10.1038/81137.
17. Kanehisa M, Furumichi M, Tanabe M. *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61. https://doi.org/10.1093/nar/gkw1092.
18. Wishart DS, Guo AC, Oler E. *et al*. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**:D622–31. https://doi.org/10.1093/nar/gkab1062.

19. Karp PD, Billington R, Caspi R. *et al*. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2019;**20**:1085–93. https://doi.org/10.1093/bib/bbx085.
20. Croft D, Mundo AF, Haw R. *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res* 2014;**42**:D472–7. https://doi.org/10.1093/nar/gkt1102.
21. Kim S, Cheng T, He S. *et al*. PubChem protein, gene, pathway, and taxonomy data collections: Bridging biology and chemistry through target-centric views of PubChem data. *J Mol Biol* 2022;**434**:167514–167514. https://doi.org/10.1016/j.jmb.2022.167514.
22. Chagoyen M, Pazos F. Tools for the functional interpretation of metabolomic experiments. *Brief Bioinform* 2013;**14**:737–44. https://doi.org/10.1093/bib/bbs055.
23. Chagoyen M, López-Ibáñez J, Pazos F. Functional analysis of metabolomics data. In: Carugo O, Eisenhaber F, (eds.), *Data Mining Techniques for the Life Sciences*, pp. 399–406. Humana, New York, NY, 2016. https://doi.org/10.1007/978-1-4939-3572-7_20.
24. Tavazoie S, Hughes JD, Campbell MJ. *et al*. Systematic determination of genetic network architecture. *Nat Genet* 1999;**22**:281–5. https://doi.org/10.1038/10343.
25. Perroud B, Lee J, Valkova N. *et al*. Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol Cancer* 2006;**5**:64. https://doi.org/10.1186/1476-4598-5-64.
26. Ethan Yixun X, Perlina A, Heather V. *et al*. Integrated pathway analysis of rat urine metabolic profiles and kidney transcriptomic profiles to elucidate the systems toxicology of model nephrotoxicants. *Chem Res Toxicol* 2008;**21**:1548–61. https://doi.org/10.1021/tx800061w.
27. Nam H, Chung BC, Kim Y. *et al*. Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. *Bioinformatics* 2009;**25**:3151–7. https://doi.org/10.1093/bioinformatics/btp558.
28. Aggio RBM, Ruggiero K, Villas-Bôas SG. Pathway activity profiling (PAPi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics* 2010;**26**:2969–76. https://doi.org/10.1093/bioinformatics/btq567.
29. Jing Gao V, Tarcea G, Karnovsky A. *et al*. Metscape: a cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* 2010;**26**:971–3.
30. Kamburov A, Cavill R, Ebbels TMD. *et al*. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 2011;**27**:2917–8. https://doi.org/10.1093/bioinformatics/btr499.
31. Misra BB, Mohapatra S. Tools and resources for metabolomics research community: a 2017–2018 update. *Electrophoresis* 2019;**40**:227–46. https://doi.org/10.1002/elps.201800428.
32. O'Shea K, Misra BB. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. *Metabolomics* **16**:2020.
33. Misra BB. New software tools, databases, and resources in metabolomics: updates from 2020. *Metabolomics* 2021;**17**:49. https://doi.org/10.1007/s11306-021-01796-1.
34. Stanstrup J, Broeckling CD, Helmus R. *et al*. The metaRbolomics toolbox in Bioconductor and beyond. *Metabolites* 2019;**9**:200. https://doi.org/10.3390/metabo9100200.
35. Yao L, Pang Z, Xia J. Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. *Brief Bioinform* 2023;**24**:bbac553–bbac553. https://doi.org/10.1093/bib/bbac553.
36. Li S, Park Y, Duraisingham S. *et al*. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* 2013;**9**:e1003123–e1003123. https://doi.org/10.1371/journal.pcbi.1003123.
37. Smith CA, Want EJ, O'Maille G. *et al*. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 2006;**78**:779–87. https://doi.org/10.1021/ac051437y.
38. Diego R H-D, Tarazona S, Martínez-Mira C. *et al*. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res* 2018;**46**:W503–9. https://doi.org/10.1093/nar/gky466.
39. Liu T, Salguero P, Petek M. *et al*. PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases. *Nucleic Acids Res* 2022;**50**:W551–9. https://doi.org/10.1093/nar/gkac352.
40. Chagoyen M, Pazos F. MBRole: enrichment analysis of metabolomic data. *Bioinformatics* 2011;**27**:730–1. https://doi.org/10.1093/bioinformatics/btr001.
41. Mak TD, Laiakis EC, Goudarzi M. *et al*. MetaboLyzer: a novel statistical workflow for analyzing postprocessed LC–MS metabolomics data. *Anal Chem* 2013;**86**:506–13. https://doi.org/10.1021/ac402477z.
42. Shen X, Wang R, Xiong X. *et al*. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun* 2019;**10**:1516. https://doi.org/10.1038/s41467-019-09550-x.
43. Pang Z, Chong J, Zhou G. *et al*. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* 2021;**49**:W388–96. https://doi.org/10.1093/nar/gkab382.
44. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 2013;**29**:1830–1. https://doi.org/10.1093/bioinformatics/btt285.
45. Costa C, Maraschin M, Rocha M. An R package for the integrated analysis of metabolomics and spectral data. *Comput Methods Programs Biomed* 2016;**129**:117–24. https://doi.org/10.1016/j.cmpb.2016.01.008.
46. Wen B, Mei Z, Zeng C. *et al*. metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinform* 2017;**18**:183. https://doi.org/10.1186/s12859-017-1579-y.
47. AlAkwaa FM, Yunits B, Huang S. *et al*. Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data. *Gigascience* 2018;**7**:giy136. https://doi.org/10.1093/gigascience/giy136.
48. Deng L, Guo F, Cheng K-K. *et al*. Identifying significant metabolic pathways using multi-block partial least-squares analysis. *J Proteome Res* 2020;**19**:1965–74. https://doi.org/10.1021/acs.jproteome.9b00793.
49. Hosseini R, Hassanpour N, Liu L-P. *et al*. Pathway-activity likelihood analysis and metabolite annotation for untargeted metabolomics using probabilistic modeling. *Metabolites* 2020;**10**:183. https://doi.org/10.3390/metabo10050183.
50. Liang D, Liu Q, Zhou K. *et al*. IP4M: An integrated platform for mass spectrometry-based metabolomics data mining. *BMC Bioinform* 2020;**21**:444. https://doi.org/10.1186/s12859-020-03786-x.
51. Ni Y, Gang Y, Chen H. *et al*. M2IA: a web server for microbiome and metabolome integrative analysis. *Bioinformatics* 2020;**36**:3493–8. https://doi.org/10.1093/bioinformatics/btaa188.
52. Deng L, Ma L, Cheng K-K. *et al*. Sparse PLS-based method for overlapping metabolite set enrichment analysis. *J Proteome*

Res 2021;**20**:3204–13. https://doi.org/10.1021/acs.jproteome.1c00064.
53. Wieder C, Lai RPJ, Ebbels TMD. Single sample pathway analysis in metabolomics: Performance evaluation and application. *BMC Bioinform* 2022;**23**:481. https://doi.org/10.1186/s12859-022-05005-1.
54. Karnovsky A, Weymouth T, Tim Hull V. *et al*. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* 2012;**28**:373–80. https://doi.org/10.1093/bioinformatics/btr661.
55. Basu S, Duren W, Evans CR. *et al*. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics* 2017;**33**:1545–53. https://doi.org/10.1093/bioinformatics/btx012.
56. Zhou G, Xia J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res* 2018;**46**:W514–22. https://doi.org/10.1093/nar/gky510.
57. Picart-Armada S, Fernández-Albert F, Vinaixa M. *et al*. FELLA: An R package to enrich metabolomics data. *BMC Bioinform* 2018;**19**:538. https://doi.org/10.1186/s12859-018-2487-5.
58. Wang Y, Liu X, Dong L. *et al*. iMSEA: a novel metabolite set enrichment analysis strategy to decipher drug interactions. *Anal Chem* 2023;**95**:6203–11. https://doi.org/10.1021/acs.analchem.2c04603.
59. Lin G, Dong L, Cheng K-K. *et al*. Differential correlations informed metabolite set enrichment analysis to decipher metabolic heterogeneity of disease. *Anal Chem* 2023;**95**:12505–13. https://doi.org/10.1021/acs.analchem.3c02246.
60. Goodacre R, Broadhurst D, Smilde AK. *et al*. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* 2007;**3**:231–41. https://doi.org/10.1007/s11306-007-0081-3.
61. Dunn WB, Ellis DI. Metabolomics: current analytical platforms and methodologies. *TrAC Trends Anal Chem* 2005;**24**:285–94.
62. Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. *J Chem* 2013;**5**:12. https://doi.org/10.1186/1758-2946-5-12.
63. Tautenhahn R, Cho K, Uritboonthai W. *et al*. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol* 2012;**30**:826–8. https://doi.org/10.1038/nbt.2348.
64. Nguyen DH, Nguyen CH, Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief Bioinform* 2019;**20**:2028–43. https://doi.org/10.1093/bib/bby066.
65. Xavier Domingo-Almenara J, Montenegro-Burke R, Guijas C. *et al*. Autonomous METLIN-guided in-source fragment annotation for untargeted metabolomics. *Anal Chem* 2019;**91**:3246–53. https://doi.org/10.1021/acs.analchem.8b03126.
66. Kuhl C, Tautenhahn R, Bottcher C. *et al*. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* 2012;**84**:283–9. https://doi.org/10.1021/ac202450g.
67. Fernández-Albert F, Llorach R, Andrés-Lacueva C. *et al*. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics* 2014;**30**:1937–9. https://doi.org/10.1093/bioinformatics/btu136.
68. Sud M, Fahy E, Cotter D. *et al*. Lmsd: Lipid maps structure database. *Nucleic Acids Res* 2007;**35**:D527–32. https://doi.org/10.1093/nar/gkl838.
69. Caspi R, Foerster H, Fulcher CA. *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2007;**36**:D623–31. https://doi.org/10.1093/nar/gkm900.
70. Wang ET, Sandberg R, Luo S. *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;**456**:470–6. https://doi.org/10.1038/nature07509.
71. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. *Front Physiol* 2015;**6**:383.
72. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012;**8**:e1002375. https://doi.org/10.1371/journal.pcbi.1002375.
73. Stouffer SA, Suchman EA, DeVinney LC. *et al*. The American soldier: adjustment during army life. In: Stouffer SA, Hovland CI, Lumsdaine AA, Sheffield FD, (eds.), *Studies in Social Psychology in World War II*, Vol. **1**. Princeton: Princeton University Press, 1949.
74. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL, (eds.), *Breakthroughs in Statistics: Methodology and Distribution*, vol. **18**, pages 66–70. Springer, New York, NY, 1970. https://doi.org/10.1007/978-1-4612-4380-9_6.
75. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;**57**:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
76. Berriz GF, King OD, Bryant B. *et al*. Characterizing gene sets with FuncAssociate. *Bioinformatics* 2003;**19**:2502–4. https://doi.org/10.1093/bioinformatics/btg363.
77. Hosack DA, Dennis G, Sherman BT. *et al*. Identifying biological themes within lists of genes with ease. *Genome Biol* 2003;**4**:R70–R70. https://doi.org/10.1186/gb-2003-4-10-r70.
78. Wieder C, Frainay C, Poupin N. *et al*. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput Biol* 2021;**17**:e1009105. https://doi.org/10.1371/journal.pcbi.1009105.
79. Pavlidis P, Qin J, Arango V. *et al*. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 2004;**29**:1213–22. https://doi.org/10.1023/B:NERE.0000023608.29741.45.
80. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012;**8**:e1002375. https://doi.org/10.1371/journal.pcbi.1002375.
81. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13. https://doi.org/10.1038/nrg1272.
82. Subramanian A, Tamayo P, Mootha VK. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**:15545–50. https://doi.org/10.1073/pnas.0506580102.
83. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci* 2013;**110**:6388–93. https://doi.org/10.1073/pnas.1219651110.
84. Liquet B, De Micheaux PL, Hejblum BP. *et al*. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* 2016;**32**:35–42. https://doi.org/10.1093/bioinformatics/btv535.
85. Jacob L, Obozinski G, Vert J-P. Group lasso with overlap and graph lasso. In: *Proceedings of the 26th annual International Conference on Machine Learning*. New York, NY, Association for Computing Machinery, 2009, pp. 433–40.
86. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Series B Stat Methodology* 2010;**72**:417–73. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

87. Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: *International Conference on Artificial Neural Networks*. Berlin, Heidelberg, Springer, 1997, pages 583–8. https://doi.org/10.1007/BFb0020217.
88. Adjaye J, Huntriss J, Herwig R. *et al*. Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* 2005;**23**:1514–25. https://doi.org/10.1634/stemcells.2005-0113.
89. Luo W, Friedman MS, Shedden K. *et al*. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinform* 2009;**10**:161. https://doi.org/10.1186/1471-2105-10-161.
90. Ellson J, Gansner E, Koutsofios L. *et al*. Graphviz—open source graph drawing tools. In: *International Symposium on Graph Drawing*. Vienna, Austria, Springer, 2001, pages 483–4. https://doi.org/10.1007/3-540-45848-4_57.
91. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2011;**28**:112–8. https://doi.org/10.1093/bioinformatics/btr597.
92. Evan Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 2007;**8**:118–27. https://doi.org/10.1093/biostatistics/kxj037.
93. Nguyen H, Shrestha S, Tran D. *et al*. A comprehensive survey of tools and software for active subnetwork identification. *Front Genet* 2019;**10**:155. https://doi.org/10.3389/fgene.2019.00155.
94. Mubeen S, Kodamullil AT, Hofmann-Apitius M. *et al*. On the influence of several factors on pathway enrichment analysis. *Brief Bioinform* 2022;**23**:bbac143–bbac143. https://doi.org/10.1093/bib/bbac143.
95. Draghici S, Khatri P, Tarca AL. *et al*. A systems biology approach for pathway level analysis. *Genome Res* 2007;**17**:1537–45. https://doi.org/10.1101/gr.6202607.
96. Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* 2009;**25**:211–7. https://doi.org/10.1093/bioinformatics/btn592.
97. Shannon P, Markiel A, Ozier O. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303.
98. Beecher CWW. The human metabolome. In: *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Springer, New York, NY, 2003, pages 311–9. https://doi.org/10.1007/978-1-4615-0333-0_17.
99. Ma H, Sorokin A, Mazein A. *et al*. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 2007;**3**:135. https://doi.org/10.1038/msb4100177.
100. Page L, Brin S, Motwani R. *et al*. *The Pagerank Citation Ranking: Bring Order to the Web*, Technical report. Stanford InfoLab, Stanford, California, 1999.
101. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011;**18**:507–22. https://doi.org/10.1089/cmb.2010.0265.
102. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom* 2003;**17**:166–73. https://doi.org/10.1002/cem.785.
103. Tsouka S, Masoodi M. Metabolic pathway analysis: advantages and pitfalls for the functional interpretation of metabolomics and lipidomics data. *Biomolecules* 2023;**13**:244. https://doi.org/10.3390/biom13020244.
104. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. *BMC bioinformatics* 2019;**20**:546. https://doi.org/10.1186/s12859-019-3146-1.
105. Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. *PloS One* 2018;**13**:e0191154–e0191154. https://doi.org/10.1371/journal.pone.0191154.
106. De Souza LP, Alseekh S, Brotman Y. *et al*. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Rev Proteom* 2020;**17**:243–55. https://doi.org/10.1080/14789450.2020.1766975.
107. Chen L, Wenyun L, Wang L. *et al*. Metabolite discovery through global annotation of untargeted metabolomics data. *Nat Methods* 2021;**18**:1377–85. https://doi.org/10.1038/s41592-021-01303-3.
108. Nguyen N, Huang H, Oraintara S. *et al*. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics* 2010;**26**:i659–65. https://doi.org/10.1093/bioinformatics/btq397.
109. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012;**13**:263–9. https://doi.org/10.1038/nrm3314.
110. Libiseller G, Dvorzak M, Kleb U. *et al*. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinform* 2015;**16**:118. https://doi.org/10.1186/s12859-015-0562-8.
111. McLean C, Kujawinski EB. AutoTuner: high fidelity and robust parameter selection for metabolomics data processing. *Anal Chem* 2020;**92**:5724–32. https://doi.org/10.1021/acs.analchem.9b04804.
112. Pang Z, Chong J, Li S. *et al*. MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. *Metabolites* 2020;**10**:186. https://doi.org/10.3390/metabo10050186.
113. Liu Y, Yang Y, Chen W. *et al*. DeepRTAlign: toward accurate retention time alignment for large cohort mass spectrometry data analysis. *Nat Commun* 2023;**14**:8188. https://doi.org/10.1038/s41467-023-43909-5.
114. Bilbao A, Munoz N, Kim J. *et al*. PeakDecoder enables machine learning-based metabolite annotation and accurate profiling in multidimensional mass spectrometry measurements. *Nat Commun* 2023;**14**:2461. https://doi.org/10.1038/s41467-023-37031-9.
115. Hoffmann MA, Nothias L-F, Ludwig M. *et al*. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol* 2022;**40**:411–21. https://doi.org/10.1038/s41587-021-01045-9.
116. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform* 2006;**7**:234. https://doi.org/10.1186/1471-2105-7-234.
117. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD. *et al*. Untargeted metabolomics strategies—challenges and emerging directions. *J Am Soc Mass Spectrom* 2016;**27**:1897–905. https://doi.org/10.1007/s13361-016-1469-y.
118. Nash WJ, Dunn WB. From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends Anal Chem* 2019;**120**:115324. https://doi.org/10.1016/j.trac.2018.11.022.
119. Dührkop K, Nothias L-F, Fleischauer M. *et al*. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* 2021;**39**:462–71. https://doi.org/10.1038/s41587-020-0740-8.
120. Skinnider MA, Wang F, Pasin D. *et al*. A deep generative model enables automated structure elucidation of novel

psychoactive substances. *Nat Mach Intell* 2021;**3**:973–84. https://doi.org/10.1038/s42256-021-00407-x.
121. Stravs MA, Dührkop K, Böcker S. *et al*. MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;**19**:865–70. https://doi.org/10.1038/s41592-022-01486-3.
122. Nothias L-F, Petras D, Schmid R. *et al*. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* 2020;**17**:905–8. https://doi.org/10.1038/s41592-020-0933-6.
123. Zhou Z, Luo M, Zhang H. *et al*. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun* 2022;**13**:6656. https://doi.org/10.1038/s41467-022-34537-6.
124. Vinaixa M, Schymanski EL, Neumann S. *et al*. Mass spectral databases for LC/MS-and GC/MS-based metabolomics: state of the field and future prospects. *TrAC Trends Anal Chem* 2016;**78**:23–35. https://doi.org/10.1016/j.trac.2015.09.005.
125. Singh A. Tools for metabolomics. *Nat Methods* 2020;**17**:24. https://doi.org/10.1038/s41592-019-0710-6.
126. Singh A. Annotating unknown metabolites. *Nat Methods* 2023;**20**:30.
127. Xavier Domingo-Almenara J, Rafael Montenegro-Burke H, Benton P. *et al*. Annotation: a computational solution for streamlining metabolomics analysis. *Anal Chem* 2018;**90**:480–9. https://doi.org/10.1021/acs.analchem.7b03929.
128. Huckvale ED, Powell CD, Jin H. *et al*. Benchmark dataset for training machine learning models to predict the pathway involvement of metabolites. *Metabolites* 2023;**13**:1120–1120. https://doi.org/10.3390/metabo13111120.
129. Lopez-Ibañez J, Pazos F, Chagoyen M. Predicting biological pathways of chemical compounds with a profile-inspired approach. *BMC Bioinform* 2021;**22**:320. https://doi.org/10.1186/s12859-021-04252-y.
130. Altman T, Travers M, Kothari A. *et al*. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinform* 2013;**14**:112. https://doi.org/10.1186/1471-2105-14-112.
131. Zolotovskaia MA, Tkachev VS, Guryanova AA. *et al*. OncoboxPD: Human 51 672 molecular pathways database with tools for activity calculating and visualization. *Comput Struct Biotechnol J* 2022;**20**:2280–91. https://doi.org/10.1016/j.csbj.2022.05.006.
132. Mubeen S, Hoyt CT, Gemünd A. *et al*. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front Genet* 2019;**10**:1203. https://doi.org/10.3389/fgene.2019.01203.
133. Chang S, Tong J, Zhu Y. *et al*. Network embedding in biomedical data science. *Brief Bioinform* 2020;**21**:182–97. https://doi.org/10.1093/bib/bby117.
134. Kovács IA, Luck K, Spirohn K. *et al*. Network-based prediction of protein interactions. *Nat Commun* 2019;**10**:1240. https://doi.org/10.1038/s41467-019-09177-y.
135. Stobbe MD, Houten SM, Jansen GA. *et al*. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst Biol* 2011;**5**:165. https://doi.org/10.1186/1752-0509-5-165.
136. McCormack T, Frings O, Alexeyenko A. *et al*. Statistical assessment of crosstalk enrichment between gene groups in biological networks. *PloS One* 2013;**8**:e54945. https://doi.org/10.1371/journal.pone.0054945.
137. Bokanizad B, Tagett R, Sahar Ansari B. *et al*. SPATIAL: a system-level PAThway impact AnaLysis approach. *Nucleic Acids Res* 2016;**44**:5034–44. https://doi.org/10.1093/nar/gkw429.
138. Silver M, Chen P, Li R. *et al*. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS Genet* 2013;**9**:e1003939. https://doi.org/10.1371/journal.pgen.1003939.
139. Maghsoudi Z, Nguyen H, Tavakkoli A. *et al*. A comprehensive survey of the approaches for pathway analysis using multi-omics data integration. *Brief Bioinform* 2022;**23**:bbac435–bbac435. https://doi.org/10.1093/bib/bbac435.
140. Chang L, Zhou G, Huiting O. *et al*. mGWAS-explorer: linking SNPs, genes, metabolites, and diseases for functional insights. *Metabolites* 2022;**12**:526. https://doi.org/10.3390/metabo12060526.
141. Chang L, Zhou G, Xia J. mGWAS-explorer 2.0: causal analysis and interpretation of metabolite–phenotype associations. *Metabolites* 2023;**13**:826. https://doi.org/10.3390/metabo13070826.