MetaboliticsDB: A Database of Metabolomics Analyses

M. Hasan Celik, Onurcan Ersen, Taj Saleh, Alper Dokay, Ahmet Enis Guven, and Ali Cakmak*

Abstract—Web-based metabolomics databases store relative metabolite abundance datasets measured under different physiological conditions. However, their pathway-level analysis capabilities are mostly limited to superimposing the measurements onto the pathways of the measured metabolites. Besides, none of the existing metabolomics databases offer tools to store, manage, compare, and search metabolomics analysis results. In this paper, we present MetaboliticsDB, which features a database of metabolomics analyses and a set of associated analytics tools. It enables users to store and compare their metabolomics analysis results against others to study, for instance, the progression of a disease. Moreover, MetaboliticsDB implements a genome-scale metabolic network-based analysis tool (i.e., Metabolitics) that performs network-based flux analysis. Besides, MetaboliticsDB features an advanced querying interface offering flexible criteria, such as listing all analyses where a certain pathway experiences a major increase in activity, to help researchers identify conditions sharing a similar mechanism. Finally, MetaboliticsDB employs AI-based models to associate the studied metabolomics data with diseases. Currently, the database contains analysis results for 2,174 individuals and 40 diseases. We demonstrate MetaboliticsDB's usage with a case study on Hepatocellular Carcinoma. Our experimental evaluation shows that MetaboliticsDB provides biologically relevant metabolic network-level analysis results, disease association with high accuracy, and a scalable architecture.
Availability: MetaboliticsDB is available online at https://metabolitics.itu.edu.tr.
Web interface source codes are available at https://github.com/itu-bioinformatics-database-lab/metabolitics-client-v2.
Web API source codes are available at https://github.com/itu-bioinformatics-database-lab/metabolitics-api-v3.

Index Terms—Metabolomics, Biological Databases, Data Analysis, Visualization, Machine Learning.

1 Introduction

Metabolomics aims to profile the relative abundances of metabolites in an organism [1]. It provides invaluable insights regarding physiological conditions, as the phenotype of diseases is often reflected in the metabolome of an organism. With the advancements in experimental methods (e.g., NMR, LC-MS, GC-MS, etc.), researchers are now able to measure the relative amounts of many metabolites with reasonable accuracy. The main challenge has been interpreting (i.e., analyzing) these measurements to understand the health and disease states, and accordingly, develop new diagnosis and treatment approaches. Many methods have been proposed to analyze metabolomics data at different levels of granularity (e.g., pathway-, reaction-, peak-level, etc.) ([2] [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]). Despite the large variety of metabolomics data analysis algorithms in literature, relatively few of them ([13], [14], [15], [16], [17]) have been made available to the researchers over web-based databases. The existing web-based tools fall under two general categories: (i) metabolic data resources and (ii) database-enabled metabolomics data analysis resources.

Metabolic data resources ([20], [21], [22], [23], [24], [25], [26], [27]) have been around for a relatively longer time, and they usually act as data dissemination hubs for the research community. The maintainers of such resources compile their data by (i) generating in their own labs, (ii) collecting from literature, and (iii) integrating data from multiple external sources. Such resources usually feature essential browsing, searching, and visualization capabilities. Among several, the most widely used resources in this category are KEGG [20], BioCyc [21], Reactome [22], etc. Besides, some other data resources focus on metabolites rather than pathways, such as HMDB [23], CheBI [24], PubChem [26], etc.

Database-enabled metabolomics data analysis resources ([13], [14], [15], [16], [17], [28]) usually work on data that is imported from the above metabolic data resources. In addition to the basic searching and visualization capabilities provided by the metabolic data resources, this category of tools also allows users to upload their own metabolomics data, and then the analysis results are provided to the user in different formats. One of the most comprehensive analysis resources in this category is MetaboAnalyst 6.0 [13]. Besides the statistical significance and discrimination analysis tools at the metabolite level, it also features pathway-level analysis in the form of enrichment and topology-based assessment. Moreover, MetaboAnalyst 6.0 allows integrated analysis of transcriptomics and metabolomics datasets. Metabolomics Workbench [14] is NIH's official data repository for metabolomics. It offers a wide array of statistical analysis tools at the metabolite level, ranging from ANOVA to hierarchical clustering. It also allows the creation of various forms of visualiations, such as boxplots, volcano plots, bar graphs, etc. However, it does not provide

- *M.H. Celik is with the Department of Computer Science, University of California, Irvine, CA. T. Saleh, and A. Dokay were with the Department of Computer Engineering, Marmara University, Istanbul, Turkey.*
- *O. Ersen, A. E. Guven, and A. Cakmak are with the Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey. Corresponding Author: A. Cakmak, ali.cakmak@itu.edu.tr*

any pathway-level analysis. MeltDB 2.0 [15] offers a number of features to annotate the metabolites, detect peaks, and eliminate noise in user-submitted raw data. Once the data is preprocessed, users may perform statistical significance tests, classification, and clustering. One differentiating feature of MeltDB is that it provides built-in support to form project groups, share data among group members or different project groups. MetabolomeExpress [16] offers similar capabilities to MeltDB 2.0 in the categories of raw data processing and statistical analysis of the processed data. XCMS Online [17] takes multi-omics integration one step further than MetaboAnalyst 6.0, and accommodates proteomics (in addition to transcriptomics) along with metabolomics measurements. It also features other common metabolite- and pathway-level analysis features (e.g., raw data processing, statistical analysis, pathway enrichment analysis, etc.) similar to the above tools. Caleydo [18] allows mapping omics data on pathways and nicely visualizes them so that users can see which pathways are being covered by the uploaded omics data, and to what degree their activities change based on the relative abundance change of metabolites in a metabolomics dataset, or the change in mRNA levels in a gene expression dataset. WebSpecmine [29] is a web-based metabolomics data analysis and mining tool built on the specmine R package. WebSpecmine enables users to upload datasets to process, visualize, and perform various analyses. WebSpecmine also trains machine learning models and predicts classes for future samples. 3Omics [30] is a web-based human metabolomics data analysis tool that offers visualization and analysis capabilities on the uploaded datasets with an emphasis on combining different omics data. Workflow4Metabolomics [31] offers data preprocessing with retention time alignment and peak extraction, and univariate analysis with nonparametric and parametric tests. POMAShiny [32] provides data preprocessing with missing value imputation, normalization, and outlier detection, univariate analysis with t-test, ANOVA, Mann-Whitney U-test, and Kruskal-Wallis test. It also offers clustering with the k-means algorithm and classification with the random forest algorithm. OmicsDashboard [33] provides multi-omics analysis capabilities. In particular, it offers multi-level hierarchical pathway-mapped visualizations of a multi-omics datasets along with several filtering options. MetaboLink [34] focuses on raw metabolomics data analysis which include tools for normalization, imputation, and statistical analysis. PathBank 2.0 [35] extends its earlier versions with increased number of pathways and improved pathway diagrams. It also contains pathway enrichment tools. WebGestalt [36] stands out with its high-performance multi-omics analysis and enrichment computation capabilities. IDSL.GOA [37] performs enrichment analysis in terms of Gene Ontology terms instead of custom pathways. MetExplore [38] allows curating, visualizing, and browsing metabolic networks. It also provides multi-omics analysis capabilities in the form of over-representation statistics. There are several other tools that perform functional enrichment analysis and visualization of metabolomics data, such as MBRole3 [39], RaMP-DB [40], PhenoMultiOmics [41], and PaintOmics [42]. In addition to the above discusses web tools, several others offer somewhat similar metabolomics analysis features but are available only as stand-alone desktop applications, such as Pathway Tools [19], SIMCA-P+ (Umetrics, Umea, Sweden), etc.

Even though the above database-enabled metabolomics analysis resources are quite extensive in the number and variety of the raw data processing and statistical analysis features that they offer, their pathway-level analysis capabilities are limited to superimposing measured metabolite changes onto their corresponding pathways. Moreover, all of the above resources consider each pathway independently of the metabolic network that they belong to. Even though MetaboAnalyst computes network-level measures, such as betweenness, centrality, etc., it only uses them for ranking pathways. Hence, the evaluation in the above resources is limited to only those pathways whose metabolites overlap with the user-submitted measurements to some extent.

In this paper, we present a novel database-enabled web resource, MetaboliticsDB, that performs metabolic activity analysis of user-provided metabolomics data in a holistic manner, considering interconnections between pathways with mass balances preserved. The main strength of MetaboliticsDB lies in its network-based metabolic flux analysis, which may offer advantages over standard pathway enrichment analysis. To this end, it employs a state-of-the-art systems-level algorithm, Metabolitics [3], which computes metabolic pathway activity differentiation scores in reference to healthy/control individuals based on the relative changes in metabolite abundances. As secondary benefits, MetaboliticsDB stores analysis results in its database, and users may compare current analysis results with previously stored analysis results of their own or other publicly available analysis results shared by other users. In this paper, we use the term ”results” to broadly refer to pathway-, reaction-, and metabolite-level change analysis at the metabolic network scale based on metabolite fold-change values provided as input. Furthermore, MetaboliticsDB allows users to make a comparison between different analysis methods (e.g., Metabolitics vs. pathway enrichment) on the same dataset. Users may also flexibly search the stored analysis results to list those where certain pathways experience significant activity increase/decrease. Finally, within a limited scope, MetaboliticsDB enables users to associate their metabolomics datasets with diseases based on AI models that it creates and maintains. MetaboliticsDB currently supports metabolomics datasets that contain annotated metabolites, which may come from targeted or untargeted studies as well as metabolite panels/kits (e.g., Biocrates kits).

We evaluate the features of MetaboliticsDB on a real metabolomics dataset obtained from individuals with hepatocellular carcinoma (HCC). Our results demonstrate that MetaboliticsDB provides (i) biologically relevant metabolic network-level analysis results along with markings through a metabolic graph, (ii) disease association to analysis results with high accuracy, and (iii) a scalable architecture supporting hundreds of simultaneous users. This paper is organized as follows. The next section summarizes data management, metabolomics analysis features, and the architecture of MetaboliticsDB. Then, we discuss our results from the evaluation of MetaboliticsDB on an HCC dataset as well as the performance and accuracy of different features. Finally,

we conclude with a discussion on how MetaboliticsDB may be employed in a wider scope with the proposed features.

2 METHODS

2.1 Database Model and Data Management

MetaboliticsDB integrates multiple data sources, including the Recon3D human metabolic network dataset [43], a human disease ontology dataset [44], a customized metabolite synonym mapping dataset, and a metabolomics analysis database. The data is managed using a hybrid approach combining relational and NoSQL models, with regular content updates.

2.1.1 Genome Scale Metabolic Network Data

The Recon3D dataset, encompassing 10,600 reactions, 5,835 metabolites, and 106 pathways, is stored as a compressed JSON file optimized for efficient storage and retrieval (see Fig. 1). Its schema, at the top level, contains three main properties, namely, pathways, reactions, and metabolites. Each of these properties is a JSON object, where the keys are the IDs of the components, and the values are the corresponding component objects. These properties use object references to establish component relationships, enabling efficient performance of client-side queries in web interfaces.

2.1.2 Human Disease Ontology Data

MetaboliticsDB incorporates 5,278 distinct disease terms, including their parent diseases (if applicable) and synonyms, sourced from the Human Disease Ontology (DO). Users can associate their metabolomics datasets with a specific disease in the database. The database content is periodically updated. DO was chosen over MESH due to its stronger integration with genomic resources, which aligns with plans to enable multi-omics analyses in future versions of MetaboliticsDB.

2.1.3 Metabolomics Analysis Data

MetaboliticsDB employs a relational database hosted on a PostgreSQL instance with hybrid NoSQL features, such as JSON fields, to store information on users, methods, diseases, datasets, metabolomics measurements, analysis results, and machine learning models. Currently, the database contains metabolomics data and corresponding analysis results for 2174 individuals across 40 distinct diseases. The database schema comprises seven tables, as illustrated in Figure 2:

- **User:** Stores user information and maintains a one-to-many relation with the **Analysis** table.
- **Methods:** Includes available analysis methods such as Metabolitics, Direct Pathway Mapping, and Pathway Enrichment.
- **Datasets:** Contains details of uploaded datasets.
- **MetabolomicsData:** Stores metabolomics measurements uploaded with datasets.
- **Analysis:** Records analysis results for samples, using two JSON columns to store reaction and pathway-level results, where each JSON field features pathway or reaction names as keys and computed flux scores as values.

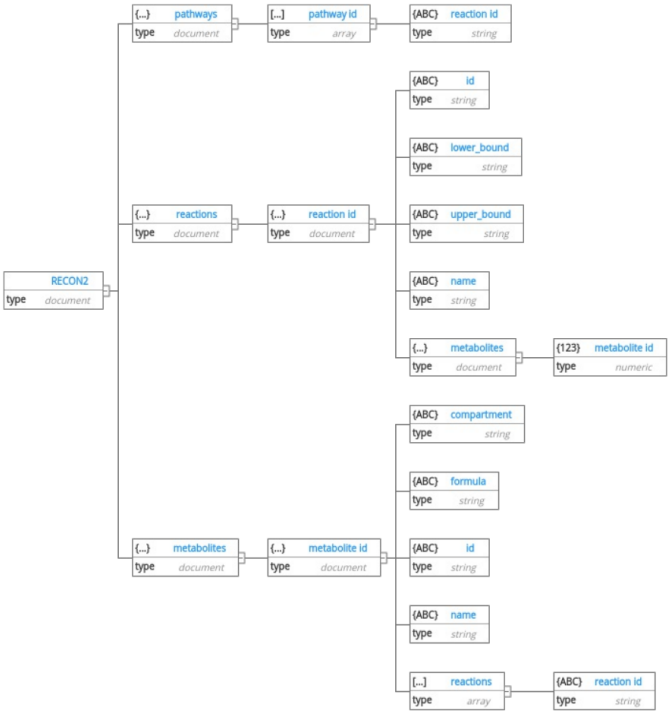

Fig. 1. The schema for the MetaboliticsDB Recon3D data (JSON)

- **Diseases:** Lists diseases known to MetaboliticsDB.
- **DiseaseModels:** Holds machine learning models trained on metabolomics analysis results.

This streamlined structure ensures efficient organization and accessibility of data for metabolomics research.

2.1.4 Multi-source Metabolite Synonym Mapping

User-provided metabolite names need to be mapped to MetaboliticsDB's database due to the lack of standardized naming conventions and the absence of synonym data in the original Recon3D dataset. MetaboliticsDB leverages BiGG [45] IDs from the Recon3D dataset and employs two strategies to enhance synonym mapping:

1) **Synonym Repository:** MetaboliticsDB consolidates alternative metabolite synonyms from datasets such as HMDB [23], KEGG [20], PubChem [26], and CheBI [24], creating a comprehensive synonym repository.
2) **RefMet Integration:** If a synonym match is not found in the local repository, RefMet nomenclature [46] is queried via HTTP POST requests. Any matched names are subsequently appended to MetaboliticsDB's customized synonym mapping dataset.

These approaches significantly enhance the recognition of metabolite synonyms, ensuring more robust and inclusive metabolomics data analysis.

2.2 Metabolomics Data Analysis

MetaboliticsDB facilitates the analysis of user-provided metabolomics datasets, aiding in the interpretation of observed changes. The platform supports three analysis methods: Metabolitics [3], Pathway Enrichment Analysis, and

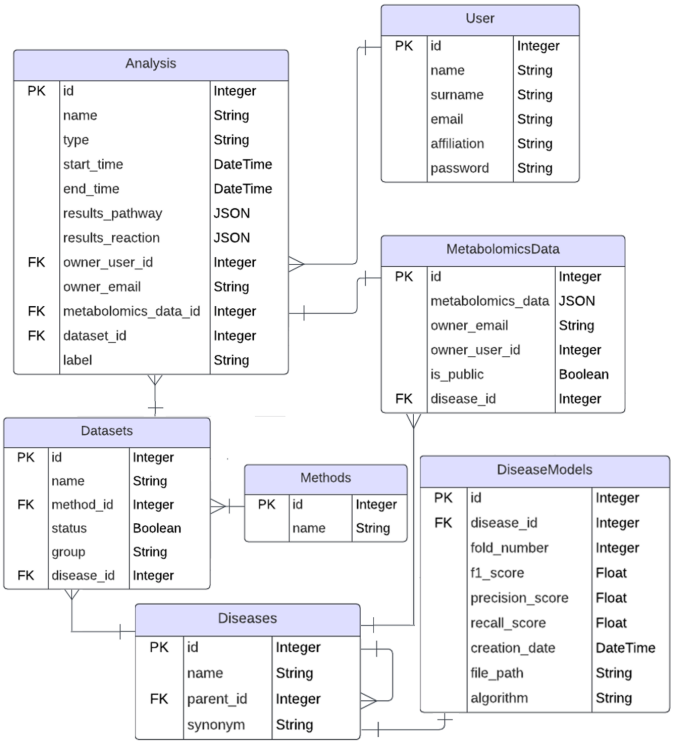

Fig. 2. Relational database schema of MetaboliticsDB

Direct Pathway Mapping. While Metabolitics provides the highest accuracy, the other methods offer faster computation and are widely used in the field. Users can leverage a built-in tool to compare the results of these methods on the same dataset. This section details the usage and principles of the analysis tools provided by MetaboliticsDB.

2.2.1 Uploading Datasets

MetaboliticsDB accepts metabolomics measurements in several file formats. Once a dataset is uploaded, fold-change values are computed relative to the average of healthy samples. The platform supports four input file formats:

- Metabolomics Workbench mwTab.
- JSON file structured as a dictionary with metabolite names as keys and abundances as values.
- CSV file containing metabolites and their abundance values.
- MetaboliticsDB's custom file format where rows are samples and columns are metabolites with cells containing the abundance value of a metabolite in a sample.

Examples of these formats are available in the online Documentation. Upon upload, MetaboliticsDB employs advanced mapping techniques to align metabolites from the input file with the genome-scale metabolic network (Recon3D) stored in its database. After mapping, a summary of mapped and unmapped metabolites is provided, allowing users to review and optionally remove any entries before analysis. The uploaded datasets stay on the server until the owners delete them.

2.2.2 Metabolic Flux Differentiation Analysis

MetaboliticsDB dynamically constructs a linear programming model [47] of the metabolic network using user-provided data. This process integrates a proprietary algorithm [3], where the total production flux for each measured metabolite is incorporated into the objective function. Fold-change values are used as coefficients for these terms.

Flux variability analysis (FVA) [48] is then performed to determine the upper and lower flux limits for each reaction. A metabolic flux differentiation score is computed for each reaction by comparing its flux boundaries to those from healthy/control samples. For each pathway, the mean differentiation score of its reactions is calculated, representing the pathway's activity deviation in individual samples. At a high level, a pathway/reaction "diff" value represents the differentiation of the pathway/reaction activity for an individual compared to healthy individuals. Hence, a positive diff value is interpreted as increased activity, while a negative diff value is interpreted as decreased activity for a pathway/reaction. These scores are stored in the database as part of the analysis.

2.2.3 Pathway Enrichment Analysis

Pathway enrichment analysis identifies pathways significantly enriched with metabolites from user-uploaded datasets. The method uses a hypergeometric distribution to compute p-values with Benjamini-Hochberg correction applied for multiple testing. MetaboliticsDB does not enforce a strict significance threshold; instead, it provides a table of computed p-values for all pathways.

2.2.4 Direct Pathway Mapping Analysis

Direct pathway mapping uses a linear model to associate metabolites from a dataset with their respective pathways. Each pathway is assigned a score based on the sum of the relative abundances of its constituent metabolites.

2.3 Tabular and Visual Analysis Results

MetaboliticsDB presents analysis results at multiple levels of detail in both visual and tabular formats. A bar plot displays the top 20 pathways, sorted by the absolute value of their diff scores (Fig. 3). Below the bar plot, all pathways and their computed diff values are listed in tabular form. Each pathway entry includes two interactive buttons:

- **Pathway Visualization:** Displays the corresponding pathway as a graph, where nodes represent metabolites and edges represent reactions. Edges are color-coded and thickened based on the reactions' diff values (see Fig. 8 for an example), enabling quick inspection of the most altered pathway regions.
- **Reaction Table:** Lists all reactions within the pathway along with their diff values in a tabular format.

All analysis results are stored in the database, accessible via the user account menu or the "Browse Analysis Results" menu. Users can mark results as "public," making them available to other users, or retain them as "private".

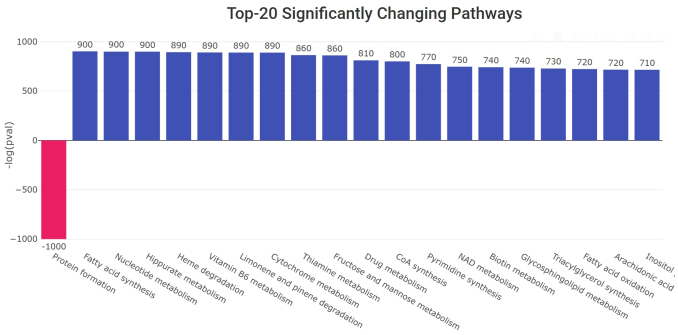

Fig. 3. Analysis results page: top-20 pathways with their diff values.

2.4 Similarity-based Disease Association

MetaboliticsDB introduces a novel feature that identifies diseases and physiological conditions most similar to a given metabolomics analysis result based on pathway-level metabolic behavior. The database stores pre-computed metabolomics analysis results for various diseases. For each user-submitted analysis, the platform identifies the top 5 diseases (e.g., diabetes) that exhibit the highest similarity to the analyzed data in terms of metabolic pathway activity distribution (see Fig. 4). To compute similarity, MetaboliticsDB represents both the user's analysis results and the disease results in the database as numeric vectors, where each value corresponds to the diff score of a pathway. Pearson correlation is then calculated between the user's vector and each disease vector. The diseases are ranked based on their correlation values, and the top 5 are presented to the user.

Similarity-based Disease Prediction

Disease	Similarity
Hepatocellular Carcinoma (Hepatoma)	1.00
Parkinson's Disease (Parkinson Disease)	0.947
Hypospadias (Familial Hypospadias)	0.931
Colon Carcinoma (Colonic Carcinoma)	0.928
Pre-Eclampsia (Preeclampsia)	0.824

Fig. 4. Similarity-based Disease Prediction

2.5 Machine Learning-based Disease Prediction

In addition to similarity-based disease association, MetaboliticsDB provides machine learning-based disease status prediction for individuals (Fig. 5). To achieve this, machine learning models are periodically trained using previously stored metabolomics data analysis results for each disease.

For each disease, five models are trained: Logistic Regression, Random Forest, Support Vector Machines, XGBoost, and an ensemble model that combines the above models using soft voting. Preprocessing steps, including feature selection and vectorization, are applied to ensure model robustness. These algorithms were selected based on their strong performance in clinical studies, as highlighted in an extensive systematic literature review [49].

Model performance is evaluated using 10-fold cross-validation, with f1-scores and standard deviations calculated to provide a comprehensive assessment. Disease datasets are split into training (90%) and testing (10%) sets, and the test set is used to evaluate model performance on unseen data. Performance results are summarized in Tables 1 and 2. The best-performing model is stored in the database as a binary file using the Python's pickle package.

These models use computed reaction metabolic differentiation scores as features. For each disease, a probability score between 0 and 1 is displayed, indicating the likelihood of having the disease.

AI-based Disease Prediction

Disease	Score
Hepatocellular Carcinoma (Hepatoma)	0.993
Stomach Cancer (Gastric Cancer)	0.929
Schizophrenia (Schizophrenia-1)	0.865
Polycystic Ovary Syndrome (PCOS)	0.708
Hypospadias (Familial Hypospadias)	0.658
Type 2 Diabetes Mellitus (Type II Diabetes Mellitus)	0.617

Fig. 5. AI-based Disease Prediction

2.6 Comparison of Analysis Results

MetaboliticsDB enables users to compare their metabolomics analysis results against (i) their own previous results and (ii) publicly available results from other users stored in the database. On the analysis results page, users can select any number of studies for comparison by checking the boxes next to them. Clicking the "compare" button at the top navigates to the comparison page.

The comparison interface includes a heatmap, where rows represent pathways with the highest variance in diff values across the selected results, and columns correspond to the chosen analysis results (Fig. 6). Each cell in the heatmap is color-coded based on the pathway's diff value, providing a clear visualization of metabolic variations.

2.7 Advanced Search Interface

A unique feature of MetaboliticsDB is the ability to search metabolomics analysis results in the database based on metabolic activity changes in specific pathways. For example, users can query analysis results where the Urea Cycle shows decreased activity, while Fatty Acid Synthesis exhibits increased activity. In addition, users can specify the magnitude of these increases or decreases for more precise filtering.

Users can dynamically add or remove pathways to construct filters with multiple conditions. These conditions are

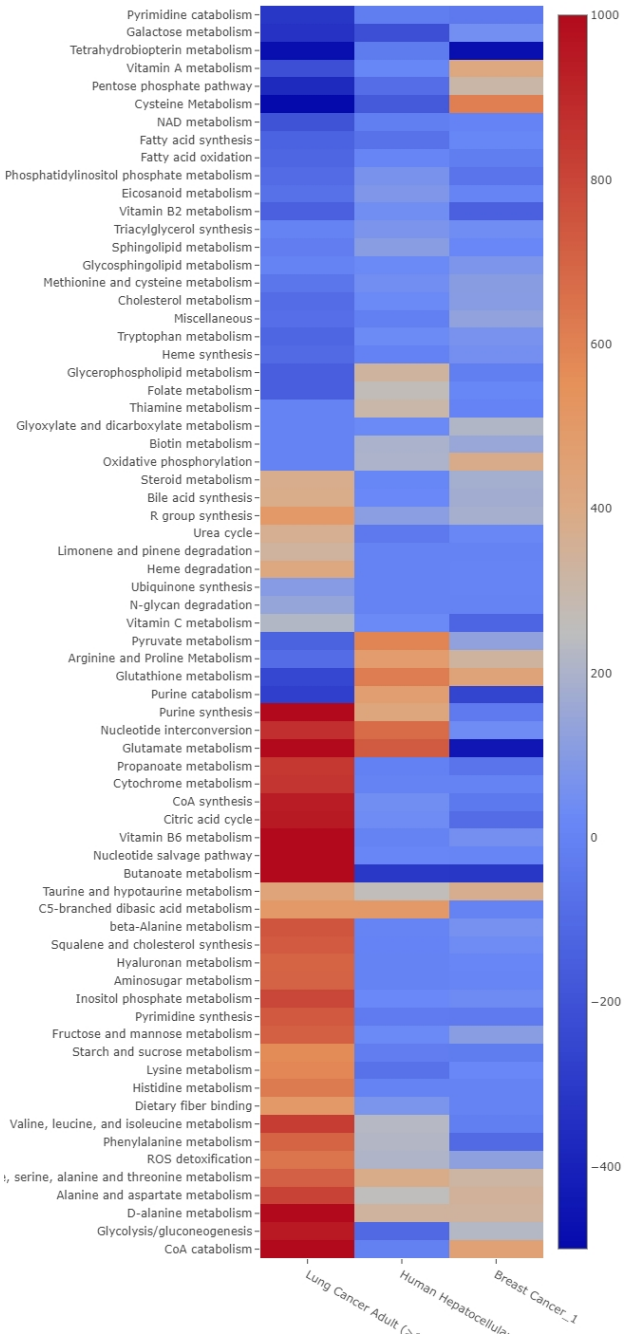

Fig. 6. MetaboliticsDB Comparison interface featuring a heatmap

evaluated using SQL "AND" semantics, ensuring that all specified criteria are satisfied simultaneously.

2.8 Browsing and Visualization Features

MetabolicticsDB includes a user-friendly search interface to locate metabolites, reactions, and pathways of interest. The search features auto-complete functionality, akin to Google Search, which dynamically suggests database entries as users type. Suggestions are categorized by entity types (e.g., metabolites, reactions) for better organization. Search results are similarly categorized, and selecting an entry redirects users to detailed information and visualizations for the selected entity.

Users can also browse the database pathway by pathway. The browsing page lists all pathways on the left-hand panel, and selecting a pathway displays its constituent reactions along with a graphical network visualization (see Fig. 8 for an example). These pathway visualizations are generated using the Escher library [50] and stored in the database.

2.9 Architecture

The architecture of MetabolicticsDB (Fig. 7) is designed to efficiently handle computational and data management requirements. The frontend is implemented using Angular, a JavaScript framework for developing sophisticated single-page web applications. Rendering and application logic are primarily executed on the client side to enhance performance and user experience.

The backend is powered by a PostgreSQL relational database, which stores analysis results and user accounts. Communication between the frontend and the database is managed via a RESTful API developed with Flask, a lightweight Python web framework. This API also provides programmatic access for researchers who wish to integrate MetabolicticsDB's functionalities into their projects. Detailed documentation for the RESTful API interfaces is available at https://metabolitics.itu.edu.tr/api/spec in OPENAPI specification.

To manage computationally intensive analysis tasks, MetabolicticsDB employs Celery, a distributed task queue, in combination with Redis for task storage. Celery workers subscribe to Redis and execute metabolomics analyses upon task submission. This approach ensures efficient handling of resource-demanding operations outside the regular HTTP request lifecycle.

Scalability is a core design consideration for MetabolicticsDB. All architectural components are dockerized, allowing distributed deployment of Celery workers across multiple instances. Additionally, the use of Angular significantly reduces server load by offloading most application logic to the client side.

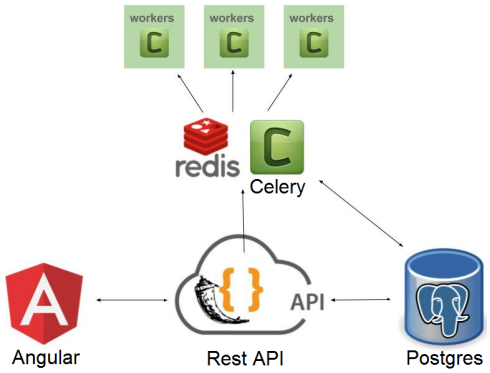

Fig. 7. Architecture overview diagram of the project.

3 Evaluation

In this section, we evaluate MetabolicticsDB and demonstrate its biological relevance through a case study on Hepatocellular Carcinoma (HCC).

3.1 A Case Study on Hepatocellular Carcinoma

To showcase MetaboliticsDB's capabilities, we analyzed a Hepatocellular Carcinoma (HCC) dataset [51] with metabolomics measurements from 177 individuals (71 healthy and 106 with liver tumors). Measurements were converted into fold changes based on healthy samples. The data was uploaded to MetaboliticsDB, analyzed using the 'Metabolitics' method, and results were displayed as bar charts and tables (available online at https://metabolitics.itu.edu.tr/panel/past-analysis/1761). Here, we discuss the top 10 pathways with the highest absolute diff scores to illustrate MetaboliticsDB's effectiveness.

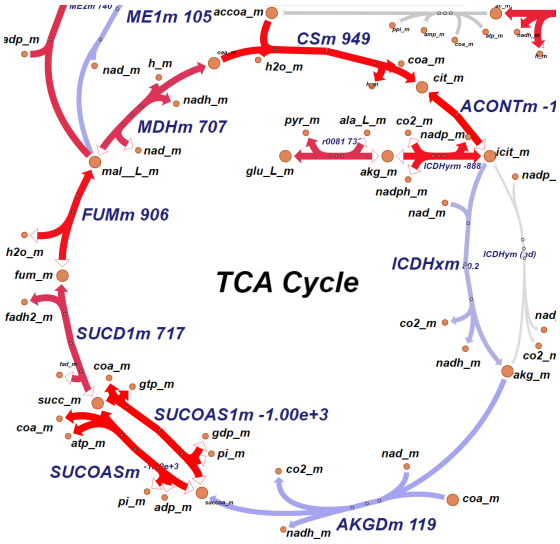

Fig. 8. Change in citric acid cycle

The protein formation pathway has the largest absolute diff value in the negative direction. PROTEIN BS reaction in the Protein formation pathway yields Torasemide-M3 metabolite. Activated metabolite Torasemide-M3 is formed from the oxidation of Torasemide [52]. Liver disease has been reported to affect the activities of Torasemide with the increased recovery of Torasemide inspected in urine [53]. This may explain the decreased activity of Torasemide activation in HCC patients.

The fatty acid synthesis pathway has the second largest diff value in the positive direction. Hepatocellular tumorigenesis has been reported to increase with abnormal activity in Fatty acid synthesis, and treatments inhibiting Fatty acid synthase enzyme might be utilized in HCC therapies [54]. Increased activity of Fatty acid synthase C180 reaction in the Fatty acid synthesis pathway observed in HCC analysis results supports this observation.

Nucleotide metabolism has the third largest diff value in the positive direction. To keep up with the fast pace of cell proliferation during tumorigenesis, increasing de novo nucleotide synthesis is essential for large-scale RNA production and DNA replication [55]. In addition, Nucleotide interconversion has a substantial diff value in the positive direction. This observation is plausible, as the essential building blocks of Nucleotide metabolism are produced by this pathway with a transformation of (d) NMP $\leftrightarrow$ (d) NDP $\leftrightarrow$ (d) NTP.

Another pathway with a large diff value in the positive direction is Hippurate metabolism. Reduced amounts of Hippurate are quantified in HCC patients due to decreased Benzoate binding proficiency [56]. Even though this is not fully aligned with our observation, the elevated quantity of Hippurate may be closely related to the diet styles of individuals [57]. Hence, more information is needed about the lifestyles of individuals to create hypotheses on this observation.

Heme degradation is another pathway with a large diff value in the positive direction. The enzyme that catalyzes heme degradation, Heme oxygenase 1 has been reported to be related to cancer progression [58]. Inhibiting the activity of Heme oxygenase 1 has been claimed to decrease HCC progression [59].

Another pathway with a large diff value in the positive direction is Vitamin B6 metabolism. The amount of Vitamin B6 compounds present in cancer cases has been reported to be less than that in control cases, and the activation levels of Pyridoxal kinase enzyme have been reported to help disease progression [60]. Increased activity of Pyridoxal kinase reaction in the Vitamin B6 metabolism seen in HCC analysis results supports these statements.

Another pathway with a large diff value in the positive direction is Limonene and pinene degradation. Limonene has been reported to inhibit the progression of HCC by suppressing cell proliferation [61]. Pinene also has been reported to inhibit cancer cell development in vitro and in vivo [62]. Decreased levels of Limonene and Pinene due to activities of Limonene and pinene degradation may be hypothesized as contributing to HCC progression.

Cytochrome metabolism is another pathway with a large diff value in the positive direction. Intrinsic clearance values indicating activity levels show an activity growth for CYP2E1, CYP2D6, and CYP2C9 cytochrome P450 types in HCC samples [63]. Increased activity of Cytochrome P450 2E1, Cytochrome P450 2C9, and Cytochrome P450 2D6 reactions in the Cytochrome metabolism pathway seen in HCC analysis results supports these observations.

Another pathway with a large diff value in the positive direction is Thiamine metabolism. The activity levels of enzymes that rely on Thiamine have been reported to increase in cancer cases [64]. Increased activity levels of Thiamine diphosphokinase, Thiamine diphosphate kinase, and Thiamine-triphosphatase reactions in the Thiamine metabolism pathway seen in HCC analysis results support these findings.

Finally, Fructose and mannose metabolism is the last among the top 10 pathways with a large diff score in the positive direction. The development of HCC was hypothesized to increase with diets rich in fructose since it enhances activity levels of the lipogenic pathway and lipid accumulation [65].

The above brief discussion illustrates that MetaboliticsDB is useful and effective in analyzing metabolomics datasets by providing insights into the underlying metabolic mechanisms. Nevertheless, we note the limitations of the literature-based evaluation due to its coarse granulation and multi-source nature of the information under different experimental settings. Thus, detailed wet lab experiments are needed for more precise conclusions.

The employed HCC dataset and the computed pathway diff scores are provided in the supplementary material.

3.1.1 Similarity-based Disease Association

MetaboliticsDB identifies diseases and physiological conditions similar to the analyzed metabolomics dataset by correlating current analysis results with previously computed disease data. To assess the relevancy, we cluster all diseases using agglomerative clustering (Pearson correlation, complete linkage) and compare it to a "ground-truth clustering" with one cluster per distinct disease, ignoring patient characteristics like gender and age. Currently, MetaboliticsDB stores 40 distinct diseases, resulting in 40 clusters. We use homogeneity and completeness metrics [66] to compare clusterings. Both metrics range from 0 to 1, with 1 being the best score. Our evaluation shows homogeneity and completeness scores of 0.94, indicating accurate cluster assignments.

3.1.2 Machine Learning-based Disease Association

We present the prediction performance of machine learning models in MetaboliticsDB. Using k-fold cross-validation (k = 10 or k = 5). Data is split based on unique patient identifiers to ensure no overlap between training and testing sets. Stratification maintains class distribution across training, validation, and testing sets. Table 1 summarizes precision, recall, and F1 scores for disease prediction models. Furthermore, we performed another evaluation with a hold-out dataset for each disease. To this end, each dataset is split into an unseen hold-out (10%) and a training datasets (90%). Based on k-fold cross-validation on the training dataset, the selected best model is tested on the hold-out dataset(see Table 2 for performance results). Results show high accuracy in classifying healthy individuals and patients.

Disease	Precision	Recall	F1	F1 SD	Alg.	K
Hepatocellular Carcinoma	0.88	0.90	0.88	0.07	LR	10
Colon Carcinoma	0.96	1.00	0.98	0.01	ENS	5
Breast Cancer	0.88	0.98	0.92	0.04	ENS	10
Stomach Cancer	0.93	0.99	0.96	0.02	LR	10
Ovarian Cancer	0.93	1.00	0.96	0.08	RF	10
Crohn's Disease	0.81	0.94	0.86	0.17	ENS	10
Asthma	0.98	1.00	0.99	0.03	ENS	10
Rheumatoid Arthritis	0.85	0.94	0.87	0.14	RF	10
Steatotic Liver Disease	0.77	0.90	0.78	0.15	RF	10
Type 2 Diabetes Mellitus	0.78	1.00	0.88	0.02	SVM	10
Wilson Disease	0.80	1.00	0.87	0.16	LR	5
Adult Respiratory Distress Syndrome	0.91	1.00	0.95	0.06	SVM	5
Androgenic Alopecia	0.83	1.00	0.88	0.16	LR	10
Chronic Fatigue Syndrome	0.75	0.85	0.76	0.03	ENS	10
Cystic Fibrosis	1.00	1.00	1.00	0.00	RF	5
Intermediate Coronary Syndrome	0.80	0.90	0.80	0.16	LR	5
Peanut Allergy	0.92	1.00	0.95	0.09	RF	5
Pre-eclampsia	0.60	1.00	0.73	0.13	SVM	5
Sarcoidosis	0.80	0.94	0.83	0.16	LR	10
Schizophrenia	0.89	1.00	0.94	0.01	RF	5
Alzheimer's Disease	0.75	0.81	0.71	0.11	RF	10
Cognitive Disorder	0.83	1.00	0.89	0.16	LR	5
Hypospadias	0.75	1.00	0.83	0.17	SVM	5
Melioidosis	0.63	1.00	0.75	0.14	LR	5
Polycystic Ovary Syndrome	0.77	0.96	0.85	0.04	SVM	10

TABLE 1
Average results of k-fold cross validation

Disease	Precision	Recall	F1	Alg.	K
Hepatocellular Carcinoma	1.00	0.91	0.95	LR	10
Colon Carcinoma	0.96	1.00	0.98	ENS	5
Breast Cancer	0.84	0.96	0.90	ENS	10
Stomach Cancer	0.93	1.00	0.96	LR	10
Ovarian Cancer	1.00	1.00	1.00	RF	10
Crohn's Disease	0.50	0.50	0.50	ENS	10
Asthma	1.00	1.00	1.00	ENS	10
Rheumatoid Arthritis	0.80	1.00	0.89	RF	10
Steatotic Liver Disease	0.80	1.00	0.89	RF	10
Type 2 Diabetes Mellitus	0.75	1.00	0.86	SVM	10
Wilson Disease	1.00	1.00	1.00	LR	5
Adult Respiratory Distress Syndrome	0.80	1.00	0.89	SVM	5
Androgenic Alopecia	1.00	0.33	0.50	LR	10
Chronic Fatigue Syndrome	0.70	0.70	0.70	ENS	10
Cystic Fibrosis	1.00	1.00	1.00	RF	5
Intermediate Coronary Syndrome	0.67	1.00	0.80	LR	5
Peanut Allergy	0.25	0.50	0.33	RF	5
Pre-eclampsia	0.33	1.00	0.50	SVM	5
Sarcoidosis	0.75	1.00	0.86	LR	10
Schizophrenia	0.95	0.95	0.95	RF	5
Alzheimer's Disease	0.50	1.00	0.67	RF	10
Cognitive Disorder	0.33	1.00	0.50	LR	5
Hypospadias	1.00	1.00	1.00	SVM	5
Melioidosis	0.50	1.00	0.67	LR	5
Polycystic Ovary Syndrome	0.74	1.00	0.85	SVM	10

TABLE 2
Test results on holdout dataset

3.2 Responsiveness Evaluation

We evaluate MetaboliticsDB's responsiveness with varying numbers of simultaneous users. Each user sends one request per second, totaling 100 requests during their browsing session. Table 3 reports the average response time in seconds and the percentage of successful responses. For this simulation, we used the open-source load testing tool Locust. The server configuration during these tests was a DELL R720 with 2 x XEON E5-2620v2 2.10 GHz CPU and 80 GB RAM running Linux Ubuntu.

Number of Users	Average Response Time (sec)	Success Rate (%)
10	0.28	100
100	0.36	100
1000	1.51	100

TABLE 3
Responsiveness evaluation results

3.3 Load Test

We performed additional performance tests for MetaboliticsDB's analysis feature. The running time depends on two factors: the size of the metabolic network and the number of metabolites in the input data. We evaluated several metabolic networks of different sizes from various organisms (obtained from BIGG [25]) (Table 4). To test the

Metabolic Net. BIGG Id	Num. of Reactions	Num. of Metabolites
e_coli_core	95	72
iAB_RBC_283	342	469
iRC1080	1706	2191
RECON1	3742	2766
RECON2	7785	5324

TABLE 4
Metabolic network size

effect of metabolomics data size, random metabolites were selected from each network, and random fold-change values were assigned. The number of metabolite measurements varied between 5 and 150 (incremented by 5, resulting in 30 datasets). The analysis was run with these artificial datasets on each metabolic network. Fig. 7 charts the average running time for each network.

MetaboliticsDB also offers similarity and machine learning-based disease prediction. Training, selecting, and saving the best model for 40 diseases (2,174 samples) takes 45 minutes and 42 seconds offline. The mean prediction time for AI-based disease association is 3.3 seconds.

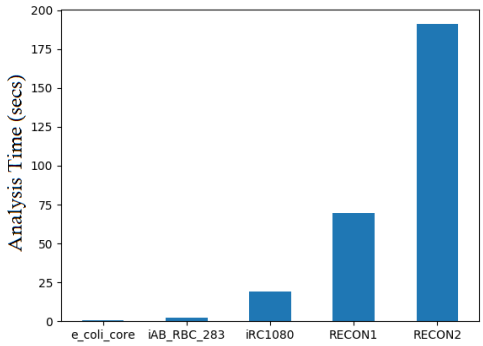

Fig. 9. Running time of MetaboliticsDB anlysis on different networks

4 Comparison

In this section, we compare MetaboliticsDB to some of the well-known tools in the field in terms of different aspects. In particular, the comparison includes MetaboAnalyst 6.0 [13], Metabolomics Workbench [14], MeltDB 2.0 [15], MetabolomeExpress [16], XCMS Online [17], Caleydo [18], WebSpecmine [29], 3Omics [30], Workflow4Metabolomics [31], and POMAShiny [32]. Even though MetaboliticsDB, is not a direct competitor of the Metabolomics Workbench, the official NIH data repository for metabolomics, and XCMS Online / MetaboAnalyst / W4M which are the references in the field of metabolomics data preprocessing/processing/annotation, we included these tools in the comparison as some of their included analysis features overlap with what MetaboliticsDB offers. Table 5 summarizes the considered aspects for comparison.

MetaboliticsDB offers metabolite name mapping, fold change scaling, and reaction diff conversion in terms of data preprocessing. Data filtration, normalization, name mapping of metabolites, and alignment and detection of peaks are some of the data preprocessing steps supported by MetaboAnalyst. Normalization and scaling are also available in Metabolomics Workbench for data preprocessing. Alignment and detection of peaks on raw data are also offered by MeltDB 2.0 and MetabolomeExpress. Data filtration is available on XCMS Online and Caleydo, however, data normalization is not supported. Data is processed with the alignment and detection of peaks steps during data upload by WebSpecmine, also other steps such as normalization and scaling are available. Normalization is supported by Workflow4Metabolomics whereas data filtration isn't available. Detection and cleaning of outliers are available on POMAShiny contrary to Workflow4Metabolomics and MetaboAnalyst 6.0.

The fold-change analysis is the univariate analysis method available in MetaboliticsDB. Volcano plots, t-tests, and fold-change analysis are among the univariate analysis methods offered by MetaboAnalyst 6.0 and MeltDB 2.0. Volcano plots and ANOVA analysis are available in Metabolomics Workbench. t-tests and fold-change analysis are provided by MetabolomeExpress, but volcano plots are not supported. Fold-change analysis is also offered by XCMS Online. ANOVA analysis is also accessible along with t-tests and fold change analysis in WebSpecmine. Non-parametric and parametric tests and ANOVA analysis are provided by Workflow4Metabolomics and POMAShiny.

MetaboliticsDB offers automatically managed machine learning classification models of type Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), and ensemble of all of the above models with soft voting (ENS). MetaboAnalyst 6.0 provides Support Vector Machine, Random Forest, and Partial Least Squares Discriminant Analysis classification methods. Random Forest and Orthogonal Partial Least Squares Discriminant Analysis classification methods are provided by Metabolomics Workbench. Support Vector Machine and Random Forest Classification methods are also supported by MeltDB 2.0. Linear Discriminant Analysis and Support Vector Machine methods are provided by WebSpecmine. Random Forest algorithm is also available in POMAShiny for classification.

Both enrichment analysis and pathway analysis are available in MetaboliticsDB. Enrichment analysis is supported by MetaboAnalyst 6.0, Metabolomics Workbench, MeltDB 2.0, XCMS Online, and 3Omics. MetaboAnalyst 6.0, MeltDB 2.0, MetabolomeExpress, XCMS Online, Caleydo, WebSpecmine, and 3Omics provide Pathway Analysis.

MetaboliticsDB, MetaboAnalyst 6.0, Metabolomics Workbench, MeltDB 2.0, XCMS Online, Caleydo, 3Omics, and POMAShiny support genome-scale metabolic networks. Python or R packages are available for MetaboliticsDB, MetaboAnalyst 6.0, Metabolomics Workbench, XCMS Online, WebSpecmine, and POMAShiny.

MetaboliticsDB, MetaboAnalyst 6.0, WebSpecmine, and POMAShiny train machine learning models and use these models for sample prediction. MetaboliticsDB periodically trains and stores predictive models on analysis results for disease prediction.

MetaboliticsDB, Metabolomics Workbench, MeltDB 2.0, MetabolomeExpress, XCMSOnline, WebSpecmine, and 3Omics enable users to compare analysis results. An advanced analysis search interface is available in MetaboliticsDB, Metabolomics Workbench, MetabolomeExpress, and XCMS Online. Metabolic flux change prediction is only available in MetaboliticsDB.

As a further comparison, we analyzed the hepatocellular carcinoma (HCC) dataset included in our case study section with the MetaboAnalyst and Metabolomics Workbench tools. MetaboAnalyst offers a network analysis feature to visualize relations between metabolites and diseases. Metabolites identified with their KEGG IDs and their fold changes in this dataset were uploaded to MetaboAnalyst Network Analysis with the Metabolite-Disease Interaction Network option. Schizophrenia was listed as the most associated

	Data Pre-processing	Univariate Analysis	Classi-fication	Enrichment Analysis	Pathway Analysis	Genome-Scale Metabolic Network Support	Python/R Package Availability	Disease Prediction	Analysis Comparison	AI Model Management	Advanced Analysis Search	Metabolic Flux Change Prediction
MetaboliticsDB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Metabo-Analyst 5.0	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗
Metabolomics Workbench	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗
MeltDB 2.0	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗
Metabolome-Express	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗
XCMS Online	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗
Caleydo	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
WebSpecmine	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗
3Omics	✗	✗	✗	✓	✓	51	✗	✗	✓	✗	✗	✗
Workflow4-Metabolomics	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
POMAShiny	✓	✓	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗

TABLE 5
Comparison of MetaboliticsDB with existing tools

disease based on the network-based scores computed by MetaboAnalyst. In contrast, MetaboliticsDB lists HCC as the most associated disease. Schizophrenia was listed as the third related disease. Furthermore, the HCC dataset was also analyzed with Pathway Enrichment Analysis features offered in both MetaboAnalyst and MetaboliticsDB. Seven pathways, namely Galactose metabolism, Methionine metabolism, Cysteine metabolism, Arginine and proline metabolism, Citric acid cycle, Alanine and aspartate metabolism, and Purine synthesis, are listed in the top-20 pathways in both tools along with high correlations outside of the top-20 list. Metabolomics Workbench lacks disease prediction and pathway analysis features. Hence, a direct comparison was not possible in these dimensions. However, it offers classification with OPLS-DA/VIP and Random Forest/VIP analyses that, given two chosen factors from the dataset, lists the most discriminating features.

5 Discussion

MetaboliticsDB stands out in several ways compared to existing tools. It offers a comprehensive data management platform for metabolomics analysis results, along with web-based tools for querying, visualizing, and studying these results at the network level. Key features include:

(i) Comparison Feature: Allows researchers to compare their datasets with known diseases or other users' public analysis results. This can reveal common mechanisms between different conditions, aiding in the sharing of therapies and providing insights for interpreting new cases. It can also help understand disease sub-types, progression stages, and drug effects through before-and-after comparisons. For example, Fig. 4 compares different cancers, highlighting the Warburg Effect.

(ii) Disease and Physiological Condition Association Tools: Helps clinicians make accurate and faster diagnoses by narrowing down possible conditions based on personalized metabolomics data analysis results using AI-powered prediction tools.

(iii) Advanced Search Interface: Facilitates the discovery of recurring patterns of metabolic fluctuations across different conditions in terms of pathway activity changes.

MetaboliticsDB's powerful analysis interface is central to its functionality. As demonstrated in the HCC case study, it enables holistic interpretations of metabolomics data, allowing users to explore metabolic mechanisms beyond biomarkers. This is achieved through its state-of-the-art personalized metabolic analysis algorithm [3].

Additionally, MetaboliticsDB can aid drug design research by highlighting changes in metabolic networks, suggesting potential drug targets, and explaining drug efficacy through pathway visualizations.

MetaboliticsDB is designed for easy generalization. Future support for multi-omics analysis will incorporate gene expression and proteomics data, enhancing the impact of MetaboliticsDB. This extension will provide invaluable insights into various physiological conditions by combining multi-omics datasets from the same patient.

6 Conclusion

In this paper, we present MetaboliticsDB, which incorporates a novel pathway-level metabolomics data analysis results database and a set of powerful associated tools running on this database. In particular, MetabolomicsDB allows users to analyze their metabolomics datasets with three different methods, store them in their private user area or share them with other users, compare them with known diseases in terms of the underlying metabolic mechanisms, visualize the changes in the metabolic network, perform basic and advanced search on metabolomics analysis results, and associate their datasets with different diseases (if any).

Acknowledgements

This work was in part supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) [Grant Numbers: 114E115 and 124N069] and the National Center for High-Performance Computing (UHEM) [Grant Number: 1009742021]. We would like to thank Beyza Turk for contributing some of the visualizations in MetaboliticsDB.

REFERENCES

[1] S. P. Putri, S. Yamamoto, H. Tsugawa, and E. Fukusaki, "Current metabolomics: Technological advances," *Journal of Bioscience and Bioengineering*, vol. 116, no. 1, pp. 9–16, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389172313000054

[2] A. J. Carroll, R. M. Salek, M. Arita, J. Kopka, and D. Walther, "Metabolome informatics and statistics: current state and emerging trends," *Frontiers in bioengineering and biotechnology*, vol. 4, 2016.

[3] A. Cakmak and M. H. Celik, "Personalized metabolic analysis of diseases," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 1014–1025, 2021.

[4] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "Web-based gene set analysis toolkit (webgestalt): update 2013," *Nucleic acids research*, vol. 41, no. W1, pp. W77–W83, 2013.

[5] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[6] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6388–6393, 2013.

[7] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS computational biology*, vol. 4, no. 11, p. e1000217, 2008.

[8] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, and S. Draghici, "Recent additions and improvements to the ontotools," *Nucleic Acids Research*, vol. 33, no. suppl_2, pp. W762–W765, 2005.

[9] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero, "A systems biology approach for pathway level analysis," *Genome research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[10] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2008.

[11] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[12] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, S. Ng, W. J. Gibb, N. J. Wang, S. Ziyad, F. Tong *et al.*, "Subtype and pathway specific responses to anticancer compounds in breast cancer," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. 2724–2729, 2012.

[13] Z. Pang, Y. Lu, G. Zhou, F. Hui, L. Xu, C. Viau, A. F. Spigelman, P. E. MacDonald, D. S. Wishart, S. Li *et al.*, "Metaboanalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation," *Nucleic Acids Research*, p. gkae253, 2024.

[14] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair *et al.*, "Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic acids research*, vol. 44, no. D1, pp. D463–D470, 2015.

[15] N. Kessler, H. Neuweger, A. Bonte, G. Langenkämper, K. Niehaus, T. W. Nattkemper, and A. Goesmann, "Meltdb 2.0–advances of the metabolomics software system," *Bioinformatics*, vol. 29, no. 19, pp. 2452–2459, 2013.

[16] A. J. Carroll, M. R. Badger, and A. H. Millar, "The metabolomeexpress project: enabling web-based processing, analysis and transparent dissemination of gc/ms metabolomics datasets," *BMC bioinformatics*, vol. 11, no. 1, p. 376, 2010.

[17] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, "Xcms online: a web-based platform to process untargeted metabolomic data," *Analytical chemistry*, vol. 84, no. 11, pp. 5035–5039, 2012.

[18] M. Streit, A. Lex, M. Kalkusch, K. Zatloukal, and D. Schmalstieg, "Caleydo: connecting pathways and gene expression," *Bioinformatics*, vol. 25, no. 20, pp. 2760–2761, 2009.

[19] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu *et al.*, "Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Briefings in bioinformatics*, vol. 11, no. 1, pp. 40–79, 2009.

[20] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research*, vol. 45, no. D1, pp. D353–D361, 2017.

[21] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller *et al.*, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic acids research*, vol. 44, no. D1, pp. D471–D480, 2015.

[22] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 44, no. D1, pp. D481–D487, 2015.

[23] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong *et al.*, "Hmdb 3.0—the human metabolome database in 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D801–D807, 2012.

[24] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams *et al.*, "The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D456–D463, 2012.

[25] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis, "Bigg models: A platform for integrating, standardizing and sharing genome-scale models," *Nucleic acids research*, vol. 44, no. D1, pp. D515–D522, 2015.

[26] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[27] B. Elliott, M. Kirac, A. Cakmak, G. Yavas, S. Mayes, E. Cheng, Y. Wang, C. Gupta, G. Ozsoyoglu, and Z. Meral Ozsoyoglu, "Pathcase: pathways database system," *Bioinformatics*, vol. 24, no. 21, pp. 2526–2533, 2008.

[28] A. E. Cicek, X. Qi, A. Cakmak, S. R. Johnson, X. Han, S. Alshalwi, Z. M. Ozsoyoglu, and G. Ozsoyoglu, "An online system for metabolic network analysis," *Database*, vol. 2014, p. bau091, 2014.

[29] S. Cardoso, T. Afonso, M. Maraschin, and M. Rocha, "Webspecmine: A website for metabolomics data analysis and mining," *Metabolites*, vol. 9, no. 10, 2019. [Online]. Available: https://www.mdpi.com/2218-1989/9/10/237

[30] T.-C. Kuo, T.-F. Tian, and Y. J. Tseng, "3omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data," *BMC systems biology*, vol. 7, pp. 1–15, 2013.

[31] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob *et al.*, "Workflow4metabolomics: a collaborative research infrastructure for computational metabolomics," *Bioinformatics*, vol. 31, no. 9, pp. 1493–1495, 2015.

[32] P. Castellano-Escuder, R. González-Domínguez, F. Carmona-Pontaque, C. Andrés-Lacueva, and A. Sánchez-Pla, "Pomashiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis," *PLOS Computational Biology*, vol. 17, no. 7, p. e1009148, 2021.

[33] S. Paley and P. D. Karp, "The omics dashboard for interactive exploration of metabolomics and multi-omics data," *Metabolites*, vol. 14, no. 1, p. 65, 2024.

[34] A. Mendes, J. F. Havelund, J. Lemvig, V. Schwämmle, and N. J. Færgeman, "Metabolink: a web application for streamlined processing and analysis of large-scale untargeted metabolomics data," *Bioinformatics*, vol. 40, no. 7, 2024.

[35] D. S. Wishart, R. Kruger, A. Sivakumaran, K. Harford, S. Sanford, R. Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski *et al.*, "Pathbank 2.0—the pathway database for model organism metabolomics," *Nucleic acids research*, vol. 52, no. D1, pp. D654–D662, 2024.

[36] J. M. Elizarraras, Y. Liao, Z. Shi, Q. Zhu, A. Pico, and B. Zhang, "WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics," *Nucleic Acids Research*, vol. 52, no. W1, pp. W415–W421, 05 2024. [Online]. Available: https://doi.org/10.1093/nar/gkae456

[37] P. Mahajan, O. Fiehn, and D. Barupal, "Idsl. goa: gene ontology analysis for interpreting metabolomic datasets," *Scientific Reports*, vol. 14, no. 1, p. 1299, 2024.

[38] L. Cottret, C. Frainay, M. Chazalviel, F. Cabanettes, Y. Gloaguen, E. Camenen, B. Merlet, S. Heux, J.-C. Portais, N. Poupin *et al.*, "Metexplore: collaborative edition and exploration of metabolic networks," *Nucleic acids research*, vol. 46, no. W1, pp. W495–W502, 2018.

[39] J. Lopez-Ibañez, F. Pazos, and M. Chagoyen, "Mbrole3: improved functional enrichment of chemical compounds for metabolomics data analysis," *Nucleic Acids Research*, vol. 51, no. W1, pp. W305–W309, 2023.

[40] J. Braisted, A. Patt, C. Tindall, T. Sheils, J. Neyra, K. Spencer, T. Eicher, and E. A. Mathé, "Ramp-db 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes," *Bioinformatics*, vol. 39, no. 1, p. btac726, 2023.

[41] Y. Shi, B. Xu, J. Chai, and C. Wang, "Phenomultiomics: an enzymatic reaction inferred multi-omics network visualization web server," *bioRxiv*, pp. 2024–04, 2024.

[42] T. Liu, P. Salguero, M. Petek, C. Martinez-Mira, L. Balzano-Nogueira, Ž. Ramšak, L. McIntyre, K. Gruden, S. Tarazona, and A. Conesa, "Paintomics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases," *Nucleic Acids Research*, vol. 50, no. W1, pp. W551–W559, 2022.

[43] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. Preciat Gonzalez, M. K. Aurich *et al.*, "Recon3d enables a three-dimensional view of gene variation in human metabolism," *Nature biotechnology*, vol. 36, no. 3, pp. 272–281, 2018.

[44] J. A. Baron, C. S.-B. Johnson, M. A. Schor, D. Olley, L. Nickel, V. Felix, J. B. Munro, S. M. Bello, C. Bearer, R. Lichenstein *et al.*, "The do-kb knowledgebase: a 20-year journey developing the disease open science ecosystem," *Nucleic acids research*, vol. 52, no. D1, pp. D1305–D1314, 2024.

[45] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis, "Bigg models: A platform for integrating, standardizing and sharing genome-scale models," *Nucleic acids research*, vol. 44, no. D1, pp. D515–D522, 2016.

[46] E. Fahy and S. Subramaniam, "Refmet: a reference nomenclature for metabolomics," *Nature methods*, vol. 17, no. 12, pp. 1173–1174, 2020.

[47] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[48] A. C. Müller and A. Bockmayr, "Fast thermodynamically constrained flux variability analysis," *Bioinformatics*, vol. 29, no. 7, pp. 903–909, 2013.

[49] E. M. Nwanosike, B. R. Conway, H. A. Merchant, and S. S. Hasan, "Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review," *International journal of medical informatics*, vol. 159, p. 104679, 2022.

[50] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson, "Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways," *PLoS computational biology*, vol. 11, no. 8, p. e1004321, 2015.

[51] T. Chen, G. Xie, X. Wang, J. Fan, Y. Qiu, X. Zheng, X. Qi, Y. Cao, M. Su, X. Wang *et al.*, "Serum and urine metabolite profiling reveals potential biomarkers of human hepatocellular carcinoma," *Molecular & Cellular Proteomics*, vol. 10, no. 7, pp. M110–004 945, 2011.

[52] S. Sahoo, H. S. Haraldsdóttir, R. M. Fleming, and I. Thiele, "Modeling the effects of commonly used drugs on human metabolism," *The FEBS journal*, vol. 282, no. 2, pp. 297–317, 2015.

[53] H. Knauf and E. Mutschler, "Clinical pharmacokinetics and pharmacodynamics of torasemide," *Clinical pharmacokinetics*, vol. 34, pp. 1–24, 1998.

[54] L. Che, P. Paliogiannis, A. Cigliano, M. G. Pilo, X. Chen, and D. F. Calvisi, "Pathogenetic, prognostic, and therapeutic role of fatty acid synthase in human hepatocellular carcinoma," *Frontiers in oncology*, vol. 9, p. 1412, 2019.

[55] X. Tong, F. Zhao, and C. B. Thompson, "The molecular determinants of de novo nucleotide biosynthesis in cancer cells," *Current opinion in genetics & development*, vol. 19, no. 1, pp. 32–37, 2009.

[56] K.-x. Wang, G.-h. Du, X.-m. Qin, and L. Gao, "1h-nmr-based metabolomics reveals the biomarker panel and molecular mechanism of hepatocellular carcinoma progression," *Analytical and Bioanalytical Chemistry*, vol. 414, no. 4, pp. 1525–1537, 2022.

[57] T. Pallister, M. A. Jackson, T. C. Martin, J. Zierer, A. Jennings, R. P. Mohney, A. MacGregor, C. J. Steves, A. Cassidy, T. D. Spector *et al.*, "Hippurate as a metabolomic marker of gut microbiome diversity: Modulation by diet and relationship to metabolic syndrome," *Scientific reports*, vol. 7, no. 1, p. 13670, 2017.

[58] C.-S. Park, D.-W. Eom, Y. Ahn, H. J. Jang, S. Hwang, and S.-G. Lee, "Can heme oxygenase-1 be a prognostic factor in patients with hepatocellular carcinoma?" *Medicine*, vol. 98, no. 26, 2019.

[59] G. Sass, P. Leukel, V. Schmitz, E. Raskopf, M. Ocker, D. Neureiter, M. Meissnitzer, E. Tasika, A. Tannapfel, and G. Tiegs, "Inhibition of heme oxygenase 1 expression by small interfering rna decreases orthotopic tumor growth in livers of mice," *International journal of cancer*, vol. 123, no. 6, pp. 1269–1277, 2008.

[60] L. Galluzzi, E. Vacchelli, J. Michels, P. Garcia, O. Kepp, L. Senovilla, I. Vitale, and G. Kroemer, "Effects of vitamin b6 metabolism on oncogenesis, tumor progression and therapeutic responses," *Oncogene*, vol. 32, no. 42, pp. 4995–5004, 2013.

[61] I. Kaji, M. Tatsuta, H. Iishi, M. Baba, A. Inoue, and H. Kasugai, "Inhibition by d-limonene of experimental hepatocarcinogenesis in sprague-dawley rats does not involve p21ras plasma membrane association," *International journal of cancer*, vol. 93, no. 3, pp. 441–444, 2001.

[62] W. Chen, Y. Liu, M. Li, J. Mao, L. Zhang, R. Huang, X. Jin, and L. Ye, "Anti-tumor effect of α-pinene on human hepatoma cell lines through inducing g2/m cell cycle arrest," *Journal of pharmacological sciences*, vol. 127, no. 3, pp. 332–338, 2015.

[63] J. Zhou, Q. Wen, S.-F. Li, Y.-F. Zhang, N. Gao, X. Tian, Y. Fang, J. Gao, M.-Z. Cui, X.-P. He *et al.*, "Significant change of cytochrome p450s activities in patients with hepatocellular carcinoma," *Oncotarget*, vol. 7, no. 31, p. 50612, 2016.

[64] J. A. Zastre, R. L. Sweet, B. S. Hanberry, and S. Ye, "Linking vitamin b1 with cancer cell metabolism," *Cancer & metabolism*, vol. 1, no. 1, pp. 1–14, 2013.

[65] L. Chávez-Rodríguez, A. Escobedo-Calvario, S. Salas-Silva, R. U. Miranda-Labra, L. Bucio, V. Souza, M. C. Gutiérrez-Ruiz, and L. E. Gomez-Quiroz, "Fructose consumption and hepatocellular carcinoma promotion," *Livers*, vol. 1, no. 4, pp. 250–262, 2021.

[66] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure." in *EMNLP-CoNLL*, vol. 7, 2007, pp. 410–420.