大規模メタボロミクスデータの再解析を実現する統合プラットフォームの構築

○E. Hayakawa[1,2], M. Takahashi[1], K. Nishida[3], T. Oka[3], H. Tsugawa[3], S. Kawano[4]
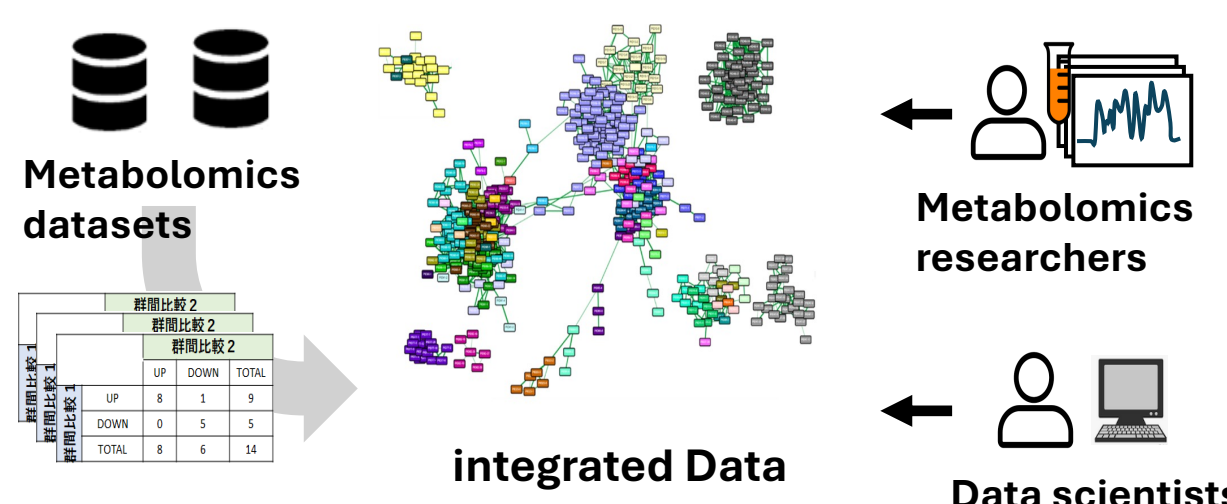
(Metabolome Informatics Research Team, Center for Sustainable Resource Science, RIKEN[1], Department of Bioscience and Bioinformatics, Kyushu Institute of Technology[2], Department of Biotechnology and Life Science, Tokyo University of Agriculture and Technology[3], School of Frontier Engineering, Kitasato University[4]
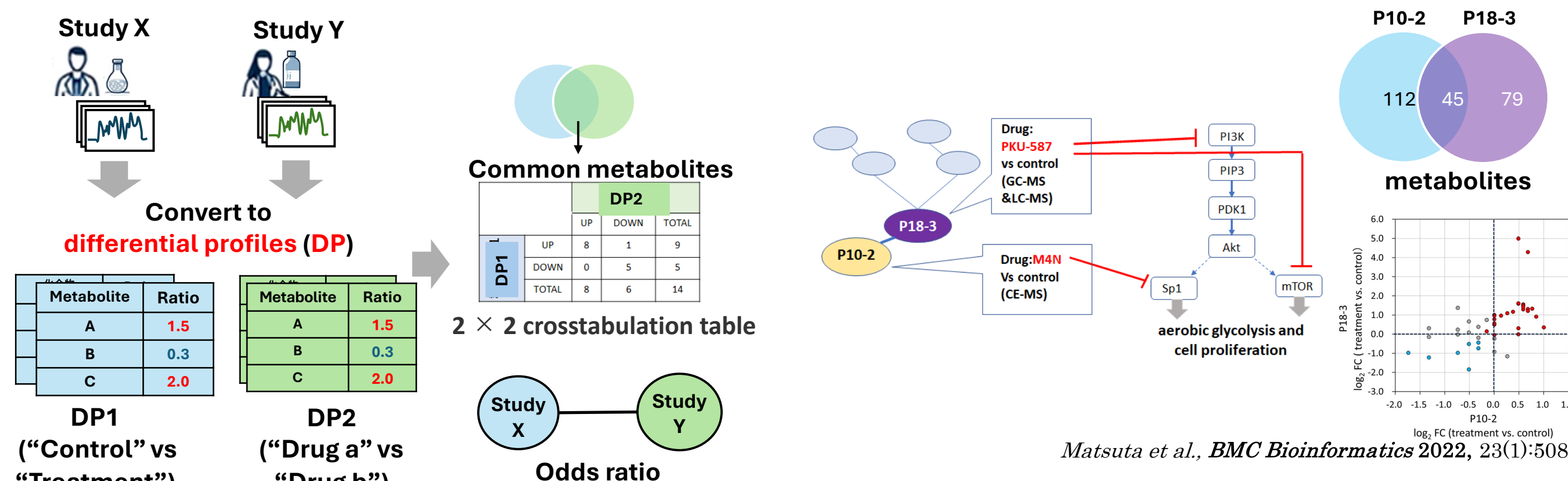
Summary

Metabolomics has advanced significantly, yielding extensive datasets stored in data repositories. However, the full potential of these data is limited by difficulties in integrating and reanalyzing information from different studies due to variations in analytical settings. We have been developing a network-based data integration platform to reanalyze cross-study metabolomics data. By quantifying similarities in differential metabolite profiles and analyzing textual metadata, we create a unified network linking metabolic data and scientific concepts. This method enhances data access and promotes cross-disciplinary scientific advancement.

Background

• **Challenges in Metabolomics Data Integration**

LC-MS inst. A
LC-MS inst. B
CE-MS

Data integration: infeasible

331 metabolites

Matsuta et al., BMC Bioinformatics, 23(1), 508 (2022).

In mass spectrometry-based metabolomics, variability in instrumentation and analytical condition poses significant challenges for standardization of metabolomics data. Differences in absolute quantification and instrument-dependent signal intensities hinder data comparability. Additionally, metabolite detection consistency is impacted by variations in instrumental settings, affecting cross-platform data reliability.

Metabolomics data repositories, such as MetaboLights and Metabolomics Workbench, contain a vast amount of data derived from diverse samples, experimental conditions, and research backgrounds. However, the metadata within these repositories is often inadequate and lacks standardization. This lack of comprehensive and uniform metadata makes the reanalysis and integration of data across networks challenging.

• **Toward Data Integration**

Metabolomics datasets
Metabolomics researchers
Data scientists
integrated Data

By addressing these challenges, we are developing a data integration platform for cross-study reanalysis.
1.**Metabolite differential profile-based Integration**: Employing a differential profile-based approach to integrate metabolomics data from various studies.
2.**Metadata-Based Integration**: Keyword-based data integration to support interpretation of complex metabolome data sets.
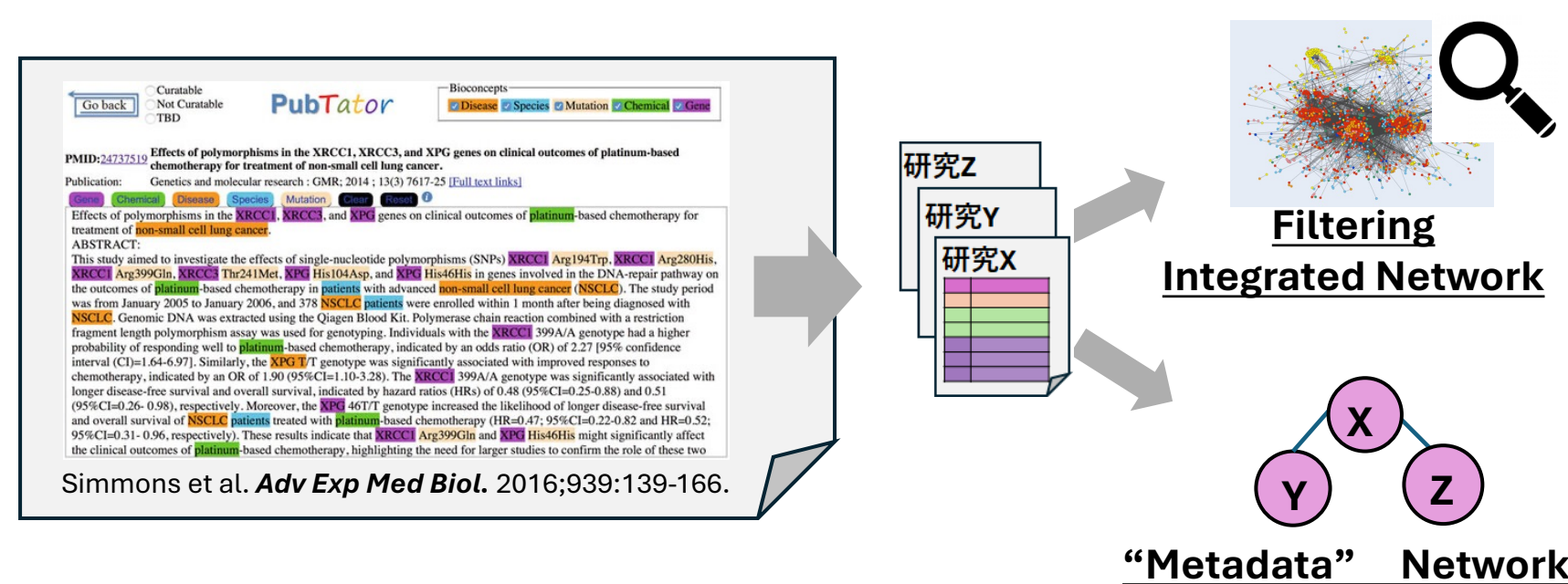3.**Unique Integrated Network for Data Exploration**: Developing an integrated network platform that facilitates the exploration of vast metabolomics data, paving the way for novel discoveries and insights.
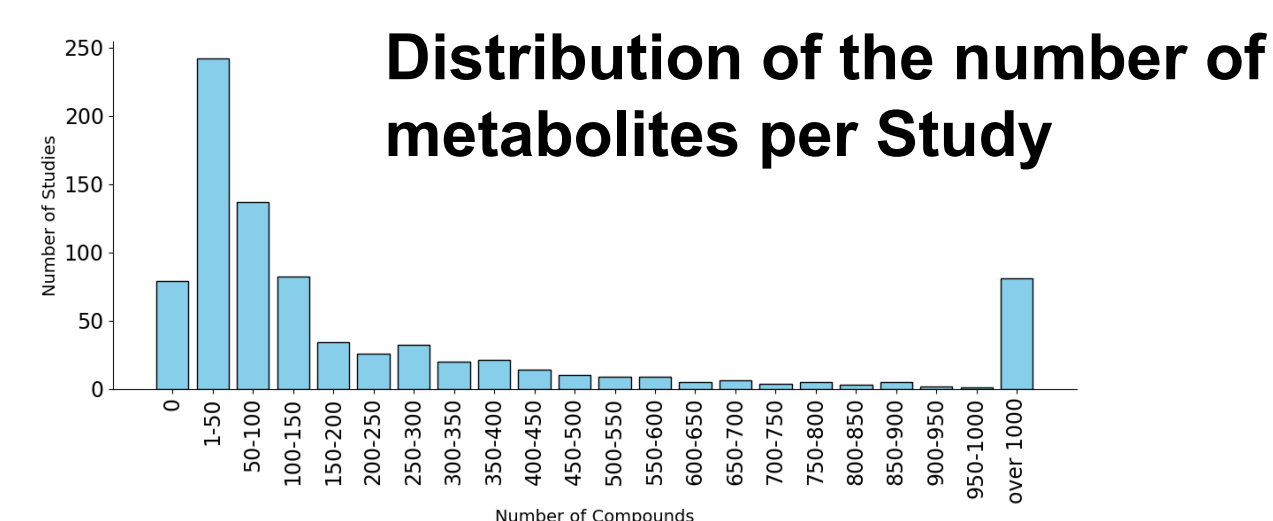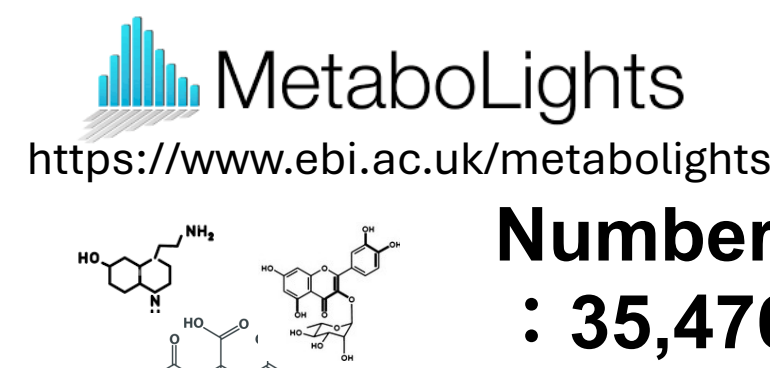
Methods

• **Differential profile-based integration**

Study X   Study Y

Convert to **differential profiles (DP)**

Metabolite	Ratio
A	1.5
B	0.3
C	2.0

DP1 ("Control" vs "Treatment")

Metabolite	Ratio
A	1.5
C	0.3
C	2.0

DP2 ("Drug a" vs "Drug b")

Common metabolites

2 × 2 crosstabulation table

Odds ratio

P10-2   P18-3

112 45 79

metabolites

aerobic glycolysis and cell proliferation

Matsuta et al., BMC Bioinformatics 2022, 23(1)-508.

We employ the iDMET approach (*Matsuta et al., BMC Bioinformatics* 23, 2022) to integrate metabolomics data from diverse analytical settings. iDMET integrates metabolomic data from different studies based on the metabolomic differential profiles (DP) between two experimental groups (e.g. Control vs Treated Sample, Drug a vs Drug b). DPs from various studies are compared using a 2x2 crosstabulation table using the number of metabolites that were up- or downregulated in a pair of differential metabolomic profiles. The odds ratio calculated based on the four numbers in the table is used as the degree of correlation between the pair. The quantified similarities are used to integrate the studies into a unified network, linking studies with similar differential profiles. This network highlights commonalities and differences in metabolic responses across studies.

• **Metadata-based Integration/Filtering**

PubTator

研究Z
研究X
研究Y

Filtering Integrated Network

"Metadata" Network

Simmons et al. *Adv Exp Med Biol.* 2016;939:139-166.

Metabolomics data repositories host diverse studies, each with unique backgrounds, making it difficult to find similarities and connections between them. We use tools to annotate key biological concepts such as diseases, species, genes, and chemicals in these studies. One such tool is PubTator (Wei et al., *Nucleic Acids Research* 47, 2019), designed for the automated annotation of biomedical literature. It uses text-mining technologies to identify and annotate key biological entities within abstracts and full-text articles.

Results and Discussion

• **Data Integration**

MetaboLights
https://www.ebi.ac.uk/metabolights
748 studies
Number of metabolites : 35,476

Distribution of the number of metabolites per Study

Number of node (Differential profile) : 68,196
Number of edges : 1,093,912 (log2 odds ratio > 4, *p*-value = 0.001)

Legend:
- Homo sapiens
- Mus musculus
- Animalia
- Rattus norvegicus
- Plantae
- Drosophila melanogaster
- Escherichia coli
- Bacteria

Metabolites most contributing network formation

Edge Count	Compound Name	Class (Classfire)
49933	aspartic acid	Aspartic acid and derivatives
40169	threonine	L-alpha-amino acids
35880	asparagine	Asparagine and derivatives
34295	methionine	Methionine and derivatives
33748	proline	Proline and derivatives
33410	tryptophan	Indolyl carboxylic acids and derivatives
32638	glutamic acid	Glutamic acid and derivatives
30994	glucose	Hexoses
29479	valine	Valine and derivatives
28591	arginine	L-alpha-amino acids
28459	phenylalanine	Phenylalanine and derivatives
28380	serine	Serine and derivatives
27349	isoleucine	Isoleucine and derivatives
27296	adenosine	Purine nucleosides
26910	ornithine	L-alpha-amino acids
26399	glutamine	Alpha amino acids
25373	lysine	D-alpha-amino acids
23238	galactitol	Sugar alcohols
22375	tyrosine	Tyrosine and derivatives
21231	hydroxy-proline	Proline and derivatives

Studies in **MetaboLights** were assessed and processed using the iDMET approach to integrate differential profiles (DPs) across studies. In this integrated network, DPs are represented as nodes, and similarities in metabolic differential changes between studies are represented as edges. The resulting network reveals a complex structure, highlighting the intricate relationships and commonalities in metabolic responses across diverse studies.

Species   Gene   Disease   Chemical

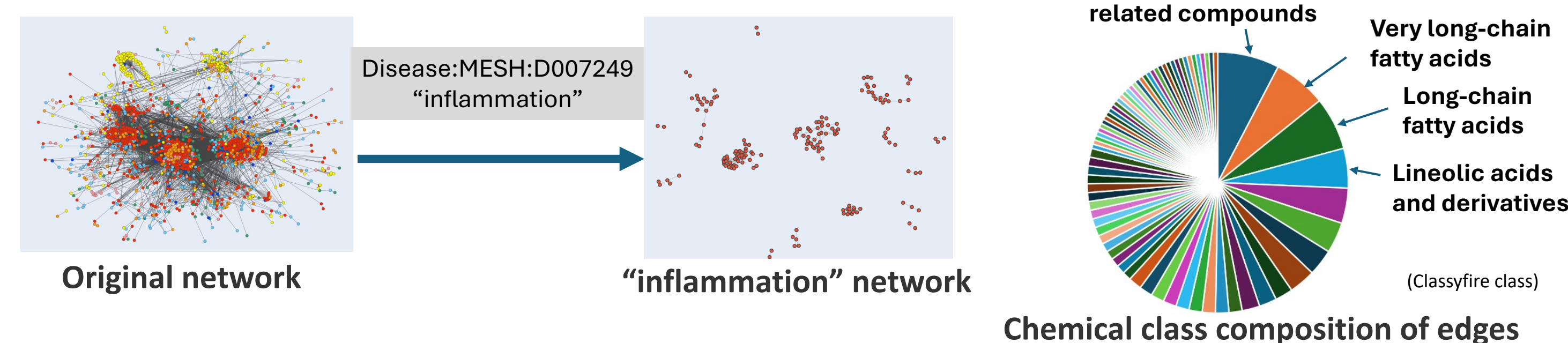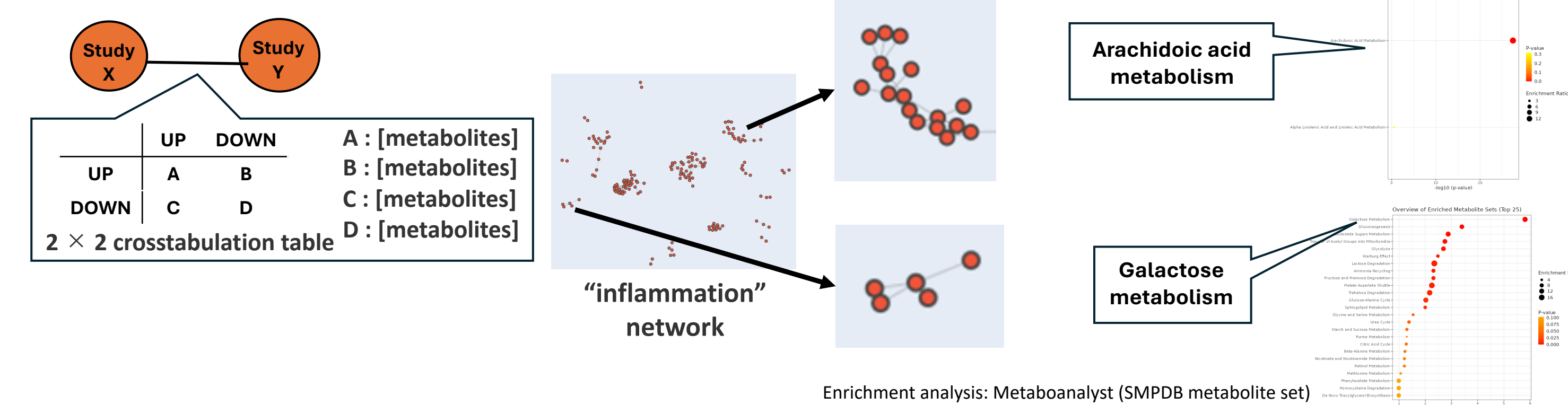338 species   198 genes   388 diseases   696 chemicals

Combined

Study descriptions were processed with PubTator to extract key biological entities such as diseases, genes, chemicals, and species. The extracted information serves as a valuable resource for assessing the data content and diversity within the repository. These biological entities were then used to evaluate meta-similarity based on the Jaccard index, facilitating a deeper understanding of the relationships and commonalities among different studies.

• **Validation and Reanalysis**

Original network
Disease:MESH:D007249 "inflammation"
"inflammation" network

Chemical class composition of edges
- Prostaglandins and related compounds
- Very long-chain fatty acids
- Long-chain fatty acids
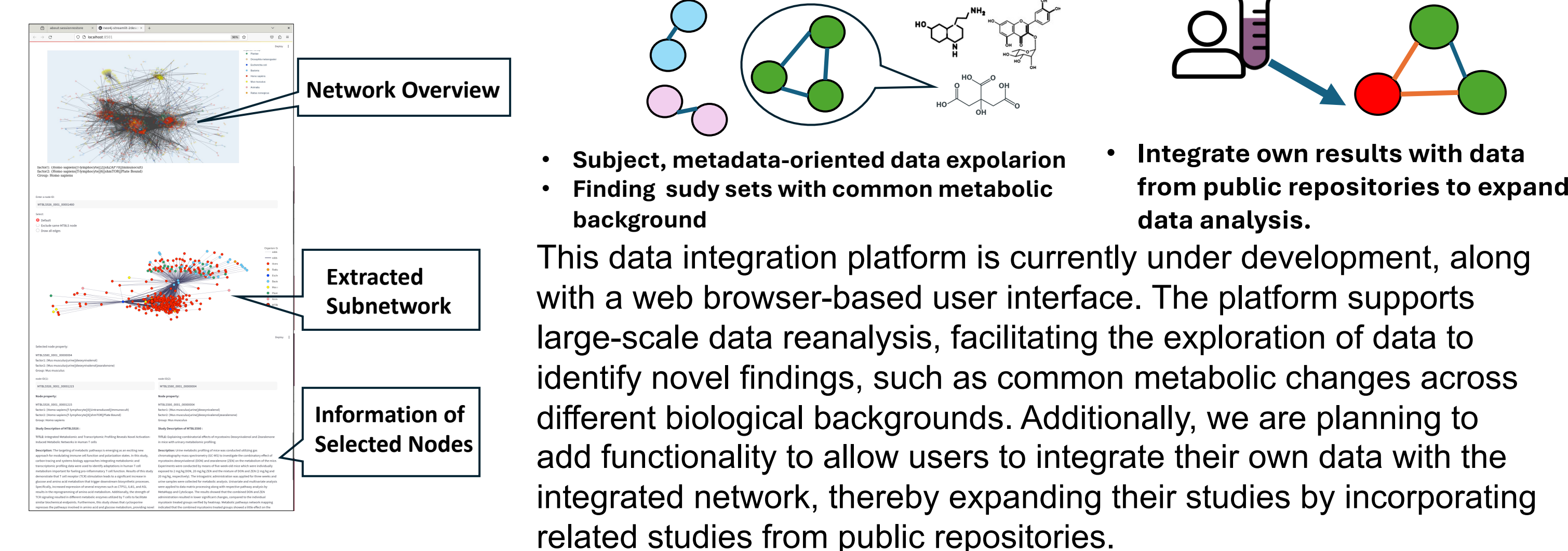- Lineolic acids and derivatives
(Classyfire class)

To validate the integrated network structure, a subnetwork related to "Inflammation" was extracted using metadata annotated with MeSH terms. The composition of metabolites involved in the edges, representing the similarity in metabolic changes, within the extracted subnetworks was analyzed. This analysis provides insights into the variable metabolites highly associated with a particular biological entity (inflammation), confirming the robustness and biological relevance of the integrated network.

Study X   Study Y

	UP	DOWN
UP	A	B
DOWN	C	D

A : [metabolites]
B : [metabolites]
C : [metabolites]
D : [metabolites]

2 × 2 crosstabulation table

"inflammation" network

Arachidoic acid metabolism
Galactose metabolism

Enrichment analysis: Metaboanalyst (SMPDB metabolite set)

The extracted "inflammation" network consists of multiple independent subnetworks, indicating that there are groups of studies with similar metabolic changes. This suggests that certain studies share common metabolic responses related to inflammation. Since the data structure of the edges contains information of metabolites contributing to the similarity of DPs, further analysis such as enrichment analysis can be applied to identify highly impacted or altered metabolic processes. The combination of the integrated network structure and metabolite-level information provides a holistic overview of the cross-study dataset, as well in-depth reanalysis of the integrated data.

• **Development as Data Integration/Reanalysis Tool**

Network Overview
Extracted Subnetwork
Information of Selected Nodes

• Subject, metadata-oriented data exploarion
• Finding sudy sets with common metabolic background

• Integrate own results with data from public repositories to expand data analysis.

This data integration platform is currently under development, along with a web browser-based user interface. The platform supports large-scale data reanalysis, facilitating the exploration of data to identify novel findings, such as common metabolic changes across different biological backgrounds. Additionally, we are planning to add functionality to allow users to integrate their own data with the integrated network, thereby expanding their studies by incorporating related studies from public repositories.

JST NBDC