

Leveraging Existing Biological Data and Making Sense of it

Finding, Accessing, Integrating and Reusing

Alberto Santos - Multi-omics Network Analytics

17 June 2025



Introduction

Multi-omics Network Analytics



Multimodal Data

Implementing tools to process, integrate, and analyse multimodal data. Diving into the benefits of harmonising multimodal data that converge to provide a comprehensive view of complex biological systems. Specifically we are interested in high-throughput multi-omics data generated using Mass spectrometry technology (proteomics and metabolomics) and metaomics data (metagenomics and metaproteomics).

Knowledge Graphs

Building High-quality Knowledge Graphs. Using and developing Knowledge Graph technologies and methods to structured data and to connect them to existing biological knowledge. These structures facilitate analysis and interpretation of complex data. We are contributing to a groundbreaking field by developing tools and methods to build, assess and investigate Knowledge Graphs and applying them to solve challenges in biology and health.

Graph Machine Learning

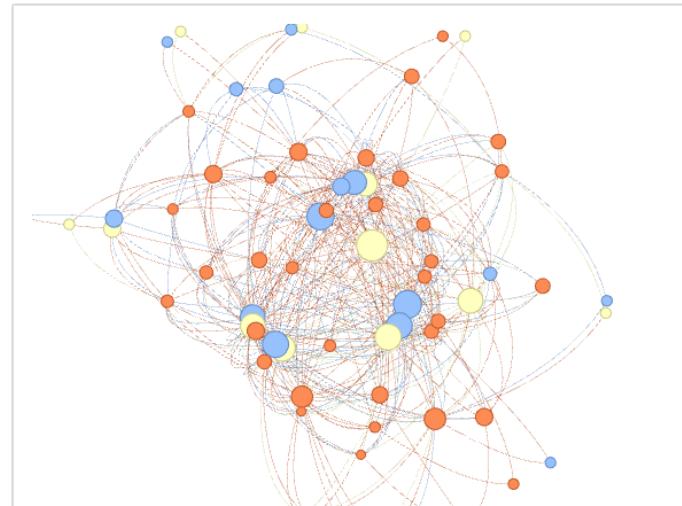
Developing and Applying Novel Methods on Graphs. Unleashing the power of Machine Learning on Graphs, a cutting-edge approach to extracting valuable insights from network data. We explore how this fusion of machine learning and graph theory helps to recognize patterns, generate predictions, and discovering new knowledge across a multitude of applications, including biological and medical networks.

Open Science

Data Science Democratisation. Focusing on data literacy training as a means to reduce inequality, and promoting open science by making all research, data content, and software open and accessible.

Microbial Communities

Exploring Microbial Communities and their Environments. Integrating multiple biological resources to unravel the assembly, interaction and adaptation mechanisms of microbial networks, offering insights into their functions and impact on ecosystems, and how changes affect those communities.



Clinical Computational Omics

Providing tools for the analysis and interpretation of clinical omics data. Integration of high-throughput omics data with computational and bioinformatics approaches to advance precision medicine and disease research. These projects aim to identify biomarkers, uncover disease mechanisms, and tailor treatments based on individual molecular profiles.

Research

A word cloud visualization representing various research topics. The most prominent words are "python", "graphs", "gnn", "knowledge-graph", "nextflow", "llm", "mans", "proteomics", and "databases". Other visible words include "deep-learning", "tools", "multi-omics", "rag", "metaomics", "geometric-deep", "pllm", "single-cell", "genomics", "machine-learning", "metagenomics", "human", "fungi", "bacteria", "nlp", and "learning". The size of each word indicates its frequency or importance, with "python" and "graphs" being the largest. The color of the words varies, with a gradient from blue to red.

python
graphs
gnn knowledge-graph
nextflow
llm mans
proteomics
databases
deep-learning tools
multi-omics rag
metaomics geometric-deep
pllm single-cell
genomics machine-learning
metagenomics human
fungi bacteria
nlp learning

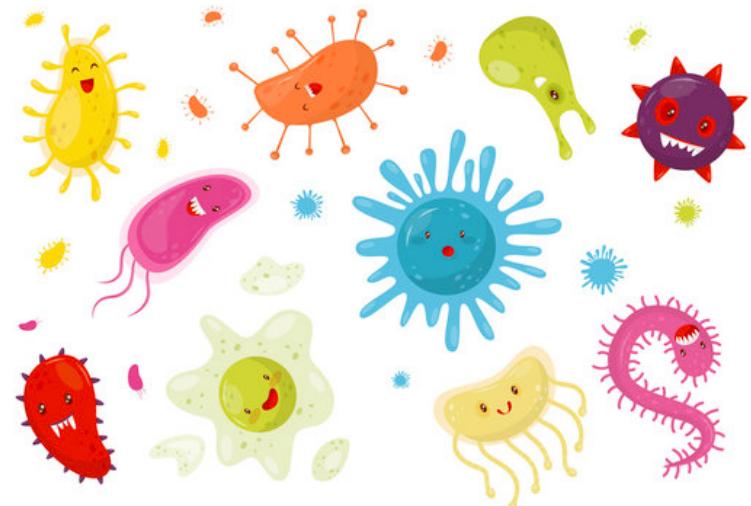
Objectives

- The **value** of **Open** and **standardised** data
- How to **generate** and **find** these data
- How to **access** and **use** them
- Some **examples**

Objectives

- The **value** of **Open** and **standardised** data
- How to **generate** and **find** these data
- How to **access** and **use** them
- Some **examples**

How to become a data parasite



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258825/>

Course website

[https://github.com/Multiomics-Analytics-Group/course synthetic biology data](https://github.com/Multiomics-Analytics-Group/course_synthetic_biology_data)

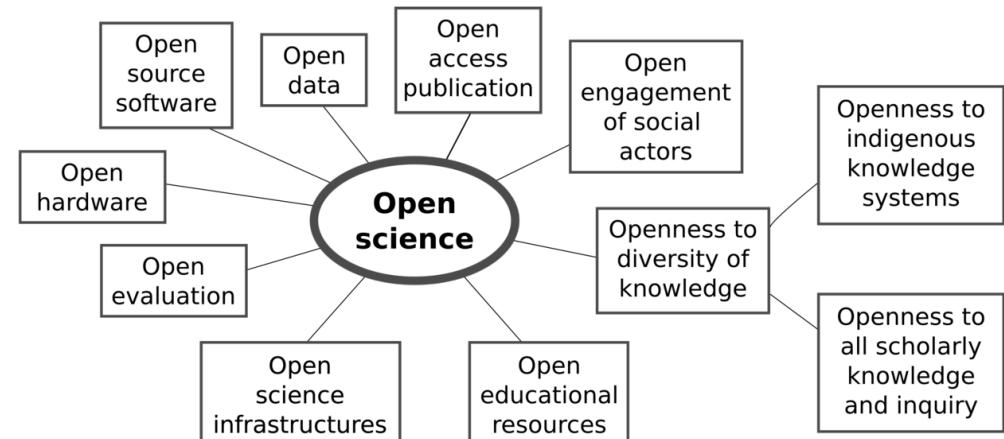


Open Science

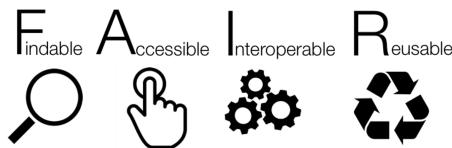
What is Open Science

Impact, Contribution, Trust

- Make scientific research **accessible** to all levels of society:
 - Publications
 - Samples
 - **Methods**
 - **Software**
 - **Data**
- Advantages:
 - **Reproducibility** and **replicability**
 - Societal **responsibility** — publicly funded, publicly available
 - **Multi-purpose** of research outputs
- Disadvantages: concerns of data **misuse**



FAIR Data and Software



- **Findable and Accessible**

- Add enough **metadata** — data about your data
- Deposit your data in **public repositories** or make them available in **databases**

[Minimum Information for Biological and Biomedical Investigations](#)

[Zenodo](#)
[Figshare](#)
[Pride](#)
[Metabolights](#)
[GEO](#)
[GitHub](#)

- Interoperable:

- Use **standard** and **open formats**
- Provide **all data needed** to reproduce your analysis

- Reusable:

- **Describe** your data well, e.g., good metadata but also
- Attach a **license**

Provide README files describing the data
Use descriptive column headers for the data tables

Challenges Sharing and Reusing

The marshmallow test – delayed gratification

- Open does not mean FAIR
- Requires an effort
- Metadata becomes the most important data
- In many cases there are no standards or multiple ones
- Most of the data out there not FAIR



<https://imgflip.com/memegenerator>

https://en.wikipedia.org/wiki/Stanford_marshmallow_experiment

Standardisation and Ontologies

- Data **standardisation** requires defining **terminologies** and **vocabularies** that:
 - Assign **unique identifiers** to entities/concepts such as proteins, genes, diseases
 - **Describe** those entities/concepts and **provide meaning**
 - **Relate** those concepts to other terms
 - Classify those entities/concepts into **categories**
- **Solution** → **Ontologies**
- **Ontology**:

formal way of representing knowledge in which concepts are described both by their meaning and their relationship to each other

A collection of terms and their definitions for a specific domain

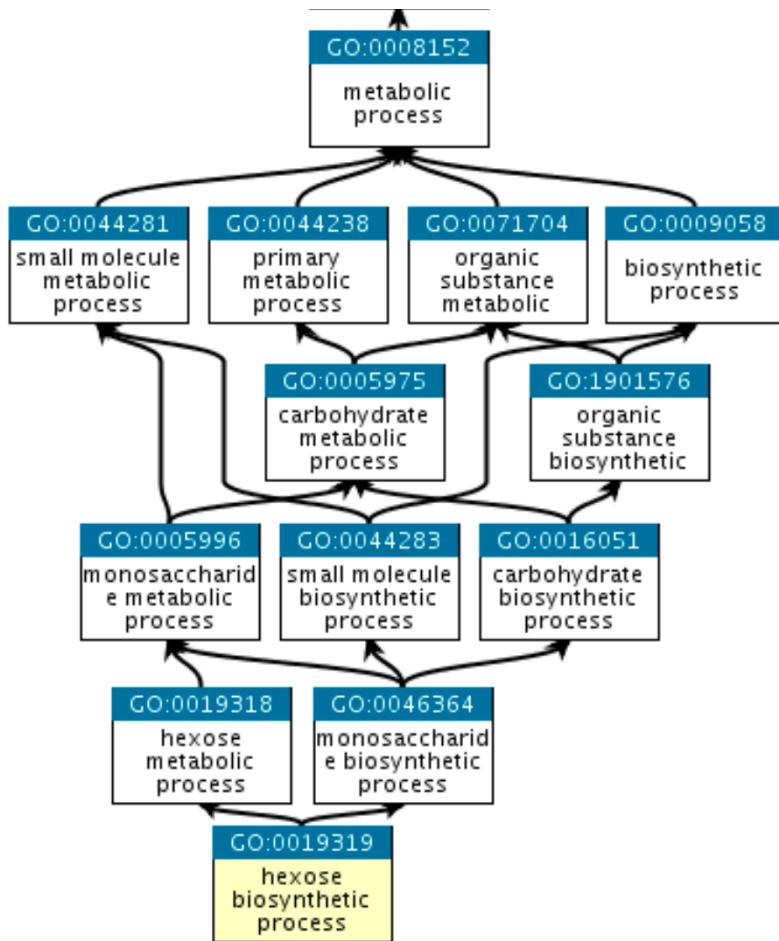


<https://www.ebi.ac.uk/ols/index>

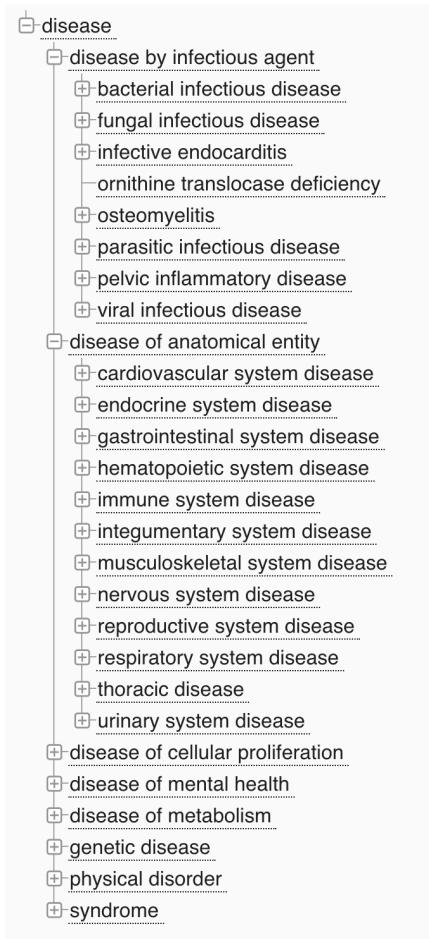
<https://www.nature.com/articles/nrg1295>

Ontologies

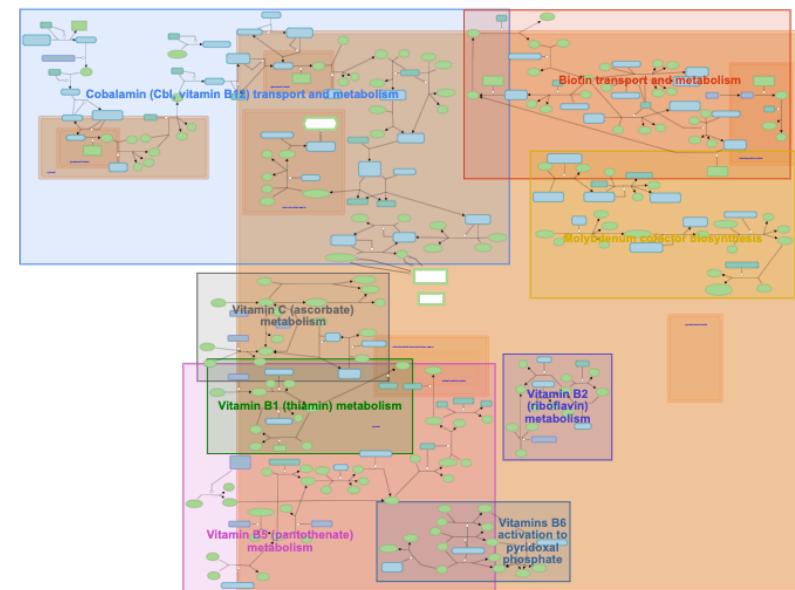
Gene Ontology



Disease Ontology



REACTOME Pathways



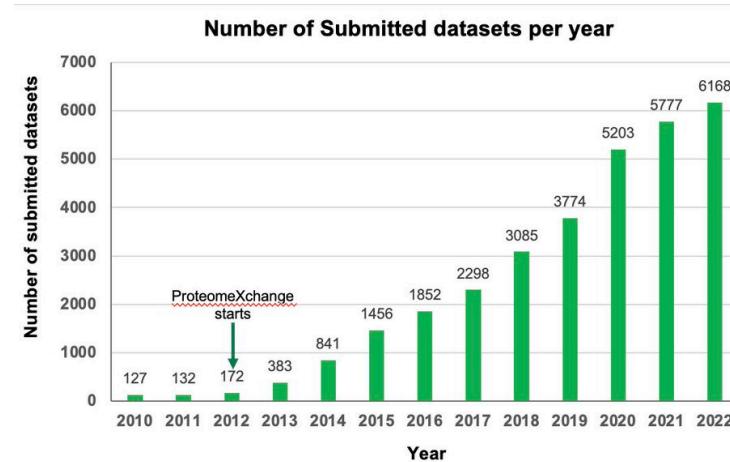
<https://www.ebi.ac.uk/ols/ontologies>

<https://reactome.org/>

<http://geneontology.org/>

Publicly Available Resources

- **Do not reinvent the wheel**
- **Extend the life and purpose** of publicly available **data**
- Build **in-silico hypotheses** before jumping into experiments (cheaper, higher success rate)
- **Download – Use – Test – Transform – Upload**
- **Growing number of resources** and **datasets** available



<https://www.ebi.ac.uk/pride/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258825/>

Data Sources

NCBI

Organisms Database

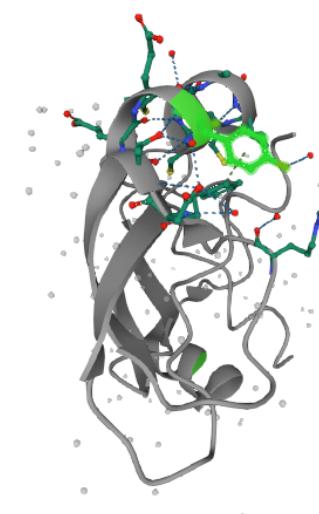
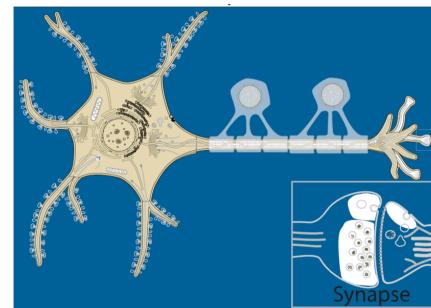
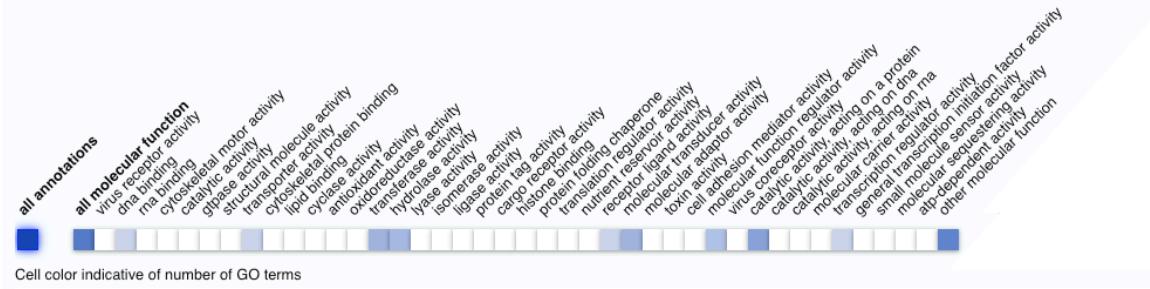
- The **NCBI Taxonomy** database allows **browsing** of the **taxonomy tree**, which contains a classification of organisms
- Provides information about:
 - Taxonomic identifier** for organisms
 - Genomic overview**
 - Link-outs to **domain-specific databases**

Entrez records			
Database name	Subtree links	Direct links	Links from type
Nucleotide	2,721,504	2,578,488	7,410
Protein	22,445,169	20,989,652	-
Structure	2,632	1,230	-
Genome	1	1	-
Popset	1,043	1,043	-
Conserved Domains	10	8	-
GEO Datasets	6,794	4,153	-
PubMed Central	111,167	111,167	-
Gene	128,094	3,304	-
SRA Experiments	41,306	33,055	-
GEO Profiles	141,600	141,600	-
Protein Clusters	7,547	7,547	-
Identical Protein Groups	3,556,802	3,524,631	-
BioProject	2,896	2,004	-
BioSample	56,942	49,071	15
Assembly	21,782	21,326	9
PubChem BioAssay	12,338	10,607	-
Taxonomy	383	1	-

UniProt

Protein Database

- Comprehensive view on **proteins**
- Aggregates information about:
 - **Function**
 - **Sequence features**
 - **Subcellular localisation**
 - **Tissue expression**
 - **Disease association**
 - **Variants and PTMs**
 - **Interactions**
 - **Structure**



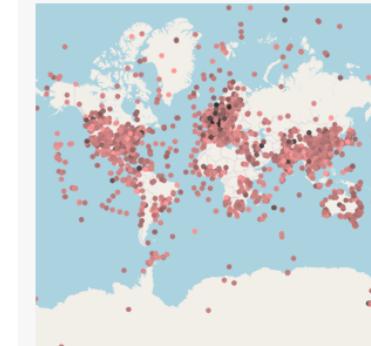
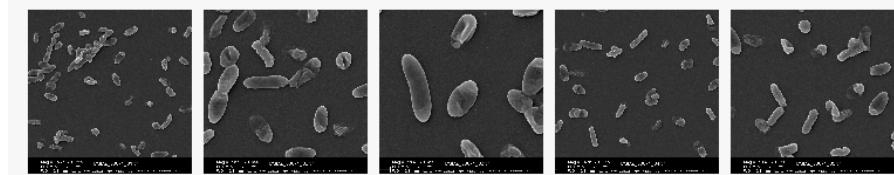
BacDive

Bacterial Database

- Largest database for **standardised bacterial phenotypic information**
- Collects information about:
 - **Bacterial classification**
 - **Morphology**
 - **Culture and growth conditions**
 - **Physiology and metabolism**
 - **Environment**
 - **Genome-based predictions**

Pseudomonas aeruginosa DSM 50071 is a thermophilic, Gram-negative, rod-shaped human pathogen of the family Pseudomonadaceae.

Gram-negative rod-shaped human pathogen thermophilic 16S sequence Bacteria genome sequence



 Aquatic Samples	5046
 Soil Samples	1669
 Animal Samples	14748
 Plant Samples	1294
Total Samples	22757

The Microbial Metabolites Database

MiMeDB

- The **human microbiome** is believed to produce or process **>55,000 different compounds** – many of which **affect** human **health, behavior and disease**
- **Microbes synthesise primary metabolites** required for their own survival, but they also **produce other compounds arising from substrates or host-derived food sources**

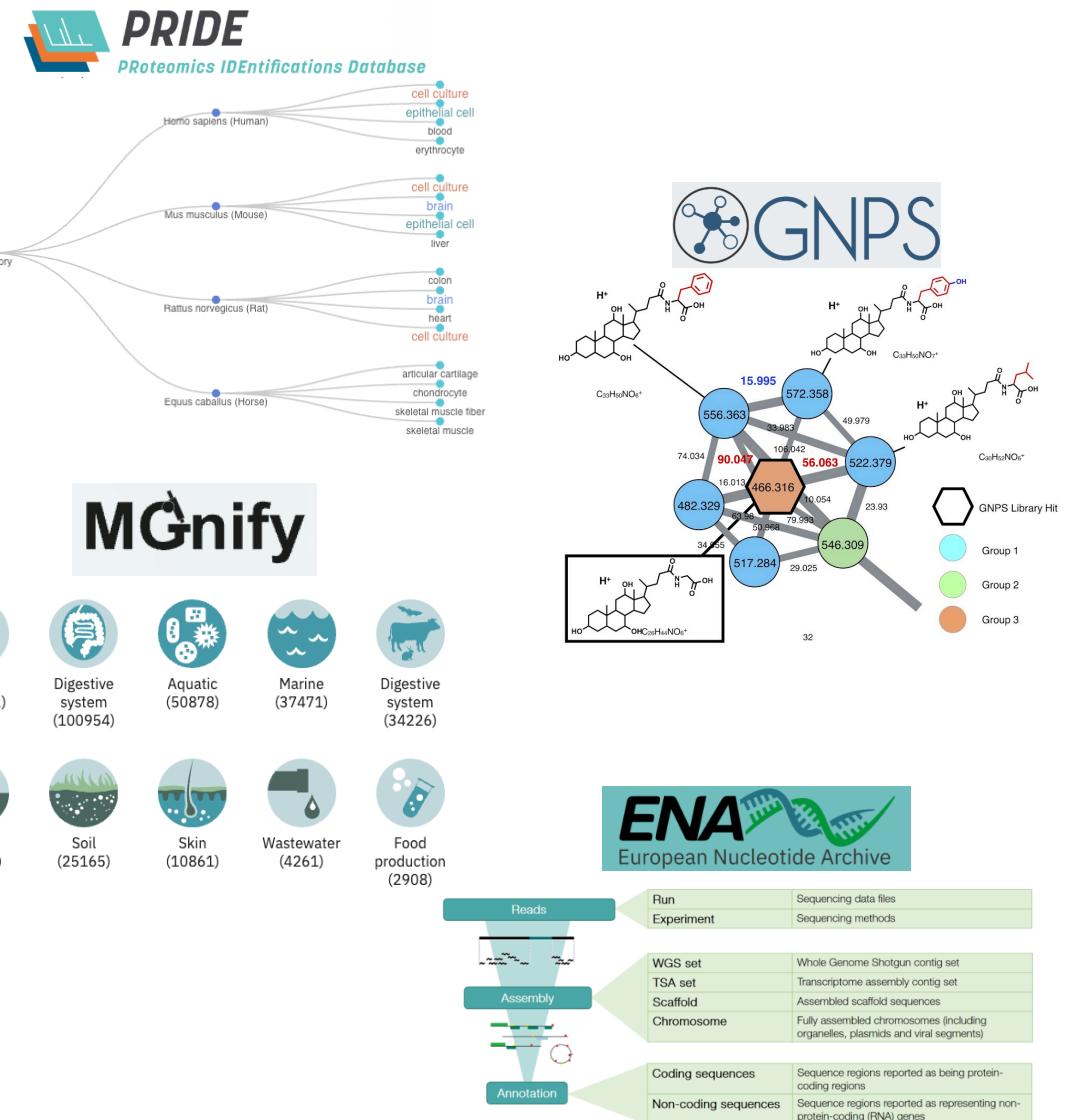
E.g., microbes transform xenobiotics from food constituents, food additives, phytochemicals, drugs, cosmetics and other exogenous or man-made chemicals

- MiMeDB is a **database of small molecule metabolites found in the human microbiome**
- Provides links between **metabolites, microbes, hosts, health and exposure** data

Meta-databases

Data Repositories

- **PRIDE:** Mass Spectrometry-based (MS-based) Proteomics
- **GNPS:** MS-based Metabolomics
- **MGnify:** Microbiome data
- **ENA:** sequencing information, covering raw sequencing data, sequence assembly information and functional annotation



Protein-Protein Interactions

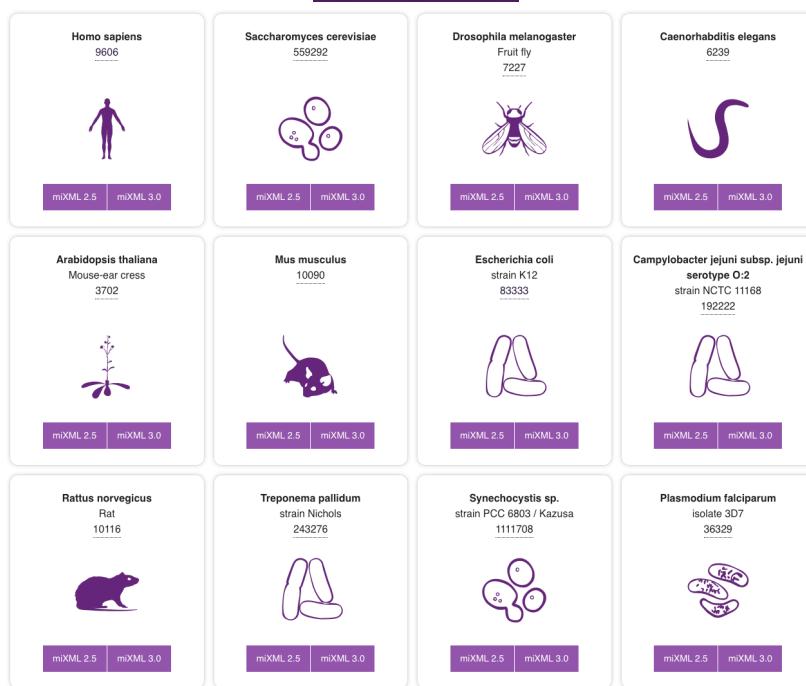
Interactions

- **STRING:** PPIs and functional enrichment



<https://string-db.org/>

- **Intact:** intra- and inter-species PPIs



<https://www.ebi.ac.uk/intact>

Other Resources

ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation <https://aledb.org/>

Web of microbes (WoM): a curated microbial exometabolomics database for linking chemistry and microbes <https://metatlas.nerc.gov/wom/project-begin.view>

MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes <http://www.liwzlab.cn/microphenodb>

MASI: microbiota—active substance interactions database <http://www.aiddlab.com/MASI/>

iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning <https://imodulondb.org/index.html>

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters <https://mibig.secondarymetabolites.org/>

FooDB: most comprehensive resource on food constituents, chemistry and biology <https://foodb.ca/>

Accessing Data

GUI vs API



Graphical User Interface

- **View** the data
- **Facilitates** initial work
- Limited to the **interface**



Application Programming Interface

- **Access** the data
- **Free** to do whatever you want
- Limited to **provided data**

Jupyter Notebooks

<https://jupyter.org/>



- **Web-based development environment** for **creating, running and sharing Python** (and other languages) code
- A **notebook** is an interactive document that combines **live code, equations, text or markdown, and visualisations** (output of your code)
- Notebooks are divided into **cells** that **run sequentially!** (Need to pay attention)
- It **requires** having **Python installed** on your local machine



Colab Notebooks

<https://research.google.com/colaboratory/faq.html>

- Google Colab is based on **Jupyter Notebook** open source project **hosted on Google's servers**
- Advantages:
 - Requires **no setup** to use (no python installation)
 - Provides **free** access to **computing resources** on Google's servers including GPUs
 - **Notebooks** can be **shared** just as you would with Google Docs or Sheets.
 - You can **import** existing **Jupyter notebooks**
- **Own data and notebooks** need to be accessed through **Google Drive** – Need Google account

Example: https://colab.research.google.com/github/biosustain/data_club/blob/main/notebooks/data_annotation/uniprot_api-solved.ipynb

Sharing Data

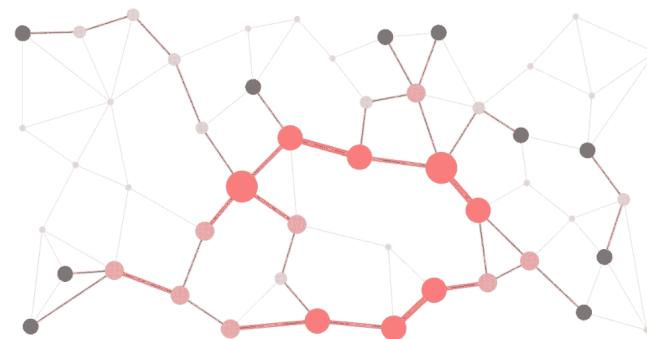
Community Repositories



scientific **data**

Working with Data

Graphs



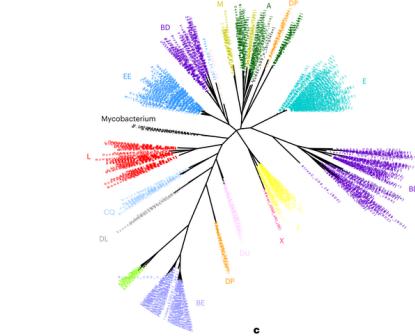
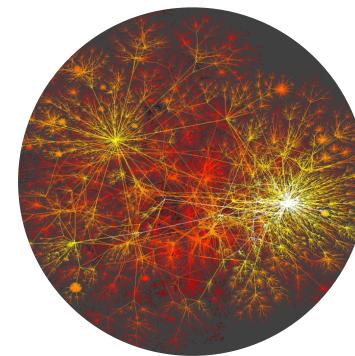
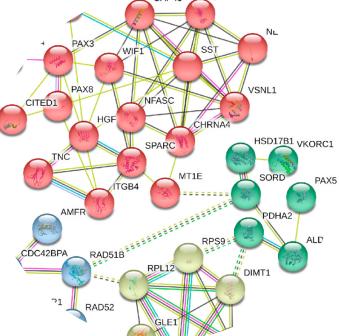
What is a Graph/Network?

- Data structures of **components (nodes)** connected by **relationships (edges)**

Social networks



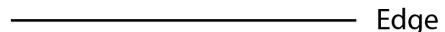
Biological networks



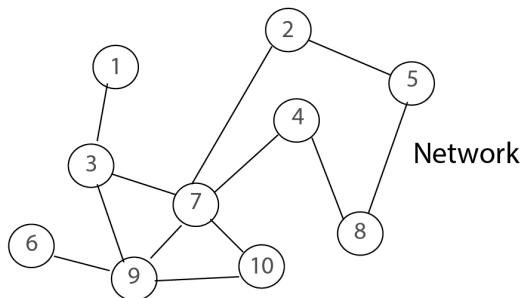
Graphs



Node



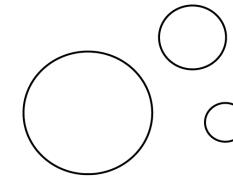
Edge



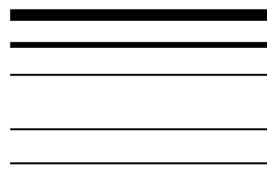
Network

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

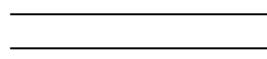
Adjacency matrix



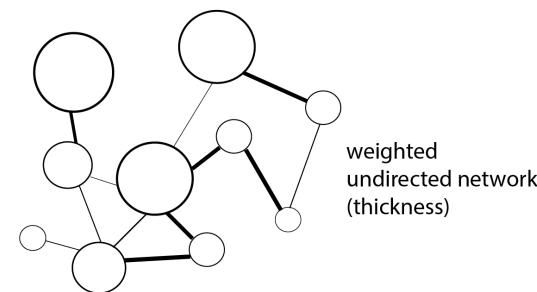
weighted nodes (size)



weighted edges (thickness)



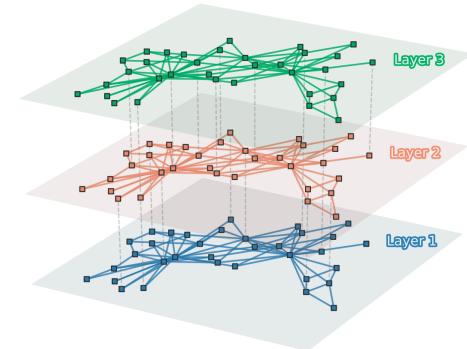
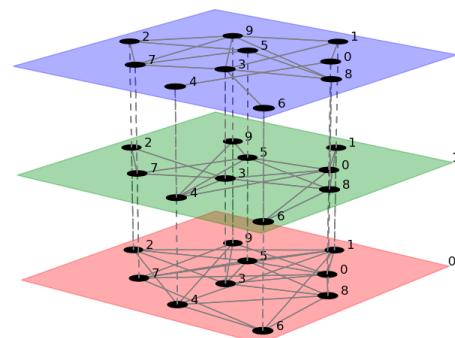
undirected edge
directed edge



weighted undirected network (thickness)

$$\begin{bmatrix} 0 & 0 & w_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_2 & 0 & w_3 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 0 & 0 & 0 & w_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & w_6 & 0 & w_7 & 0 & 0 \\ 0 & w_8 & 0 & 0 & 0 & 0 & w_9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{10} & 0 & 0 \\ 0 & w_{11} & w_{12} & 0 & 0 & 0 & 0 & w_{13} & 0 & 0 \\ 0 & 0 & 0 & w_{14} & 0 & w_{15} & 0 & 0 & w_{16} & 0 \\ 0 & 0 & w_{17} & 0 & w_{18} & 0 & 0 & 0 & w_{19} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{20} & 0 & 0 & 0 & w_{21} \end{bmatrix}$$

Weighted adjacency matrix



Layer 3

Layer 2

Layer 1

Why Graphs?

- These structures allow:
 - Quick **integration of heterogeneous data** based on relationships
 - **Graph theory** methods can be used to **analyse** and **interpret** data, e.g., topological properties can be used to explain:
 - The possible **role** of specific components
 - The **flow** of information
 - The **robustness** of the system
- **Visualize** data



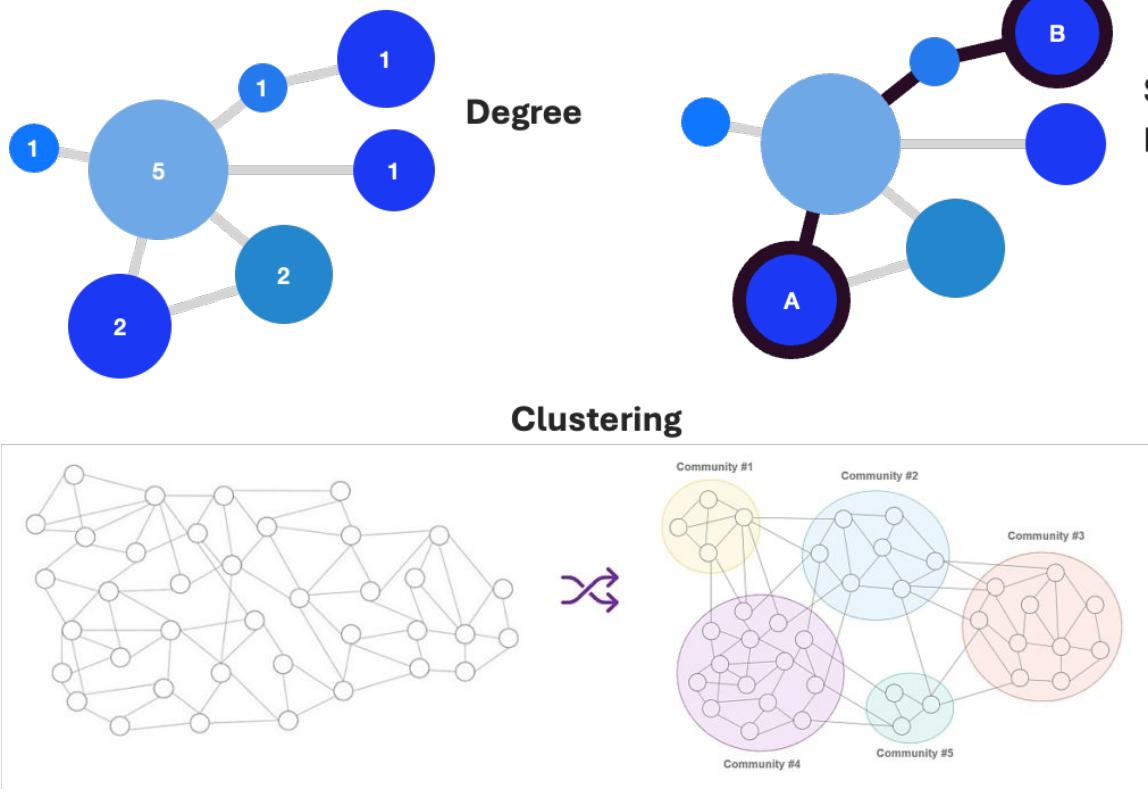
How to Analyse Graph Structures

Using and Analysing Relationships

- **Graph Theory:** algorithms that allow you to extract relevant information from the topology of the graph.
 - **Topological Features:** Centrality, degree, clustering, etc.
- **Graph Machine Learning:**
 - Embeddings
 - Graph Neural Networks

Graph Theory – Some Topological Properties

Topological properties can help extract meaningful information and identify relevant structures within the network



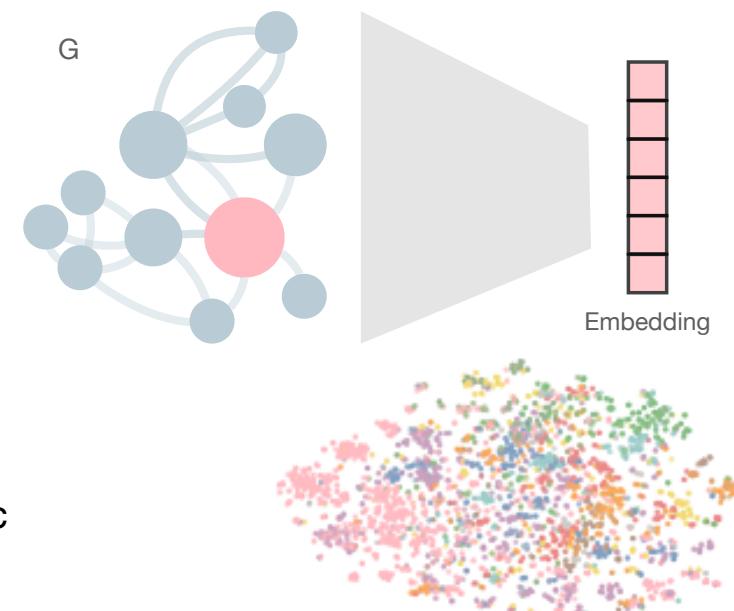
<https://timbr.ai/community-detection-algorithm/>

Graph Theory Measures and Their Application to Neurosurgical Eloquence. Cancers 2023

Embeddings

Representing graph structures

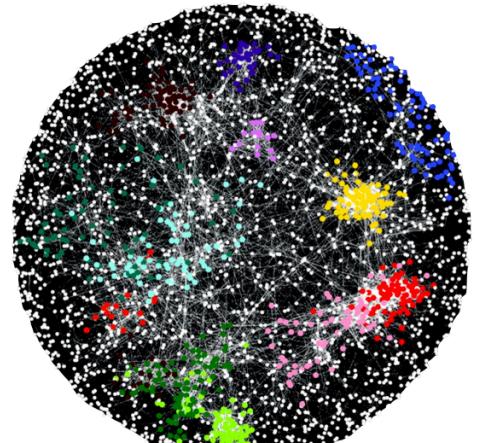
- Large biomedical networks require high computation and space —> dimensionality reduction techniques to represent nodes and edges
- Graph Embedding: graphs, nodes and edges represented as numerical features in a low dimensional space
- Use for:
Node classification
Node recommendation
Link prediction
- Methods:
 - Shallow/Walk-based approaches: Node2Vec, Path2Vec, etc.
 - Graph Neural Networks
- Python libraries: NetworkX, DGL, StellarGraph, PyTorch Geometric



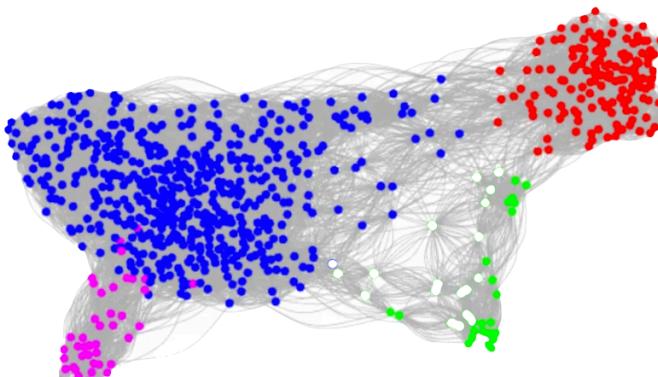
<https://www.nature.com/articles/s41467-022-33026-0>

Graphs in Biology

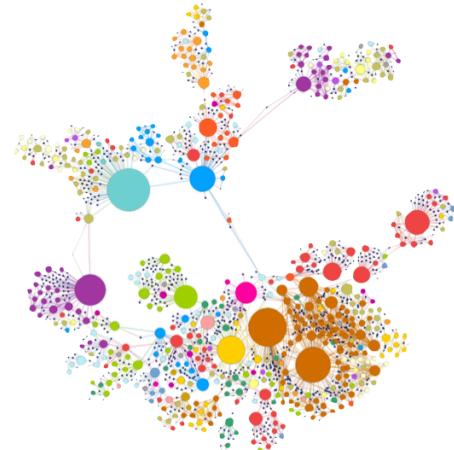
Protein-protein Interaction Networks



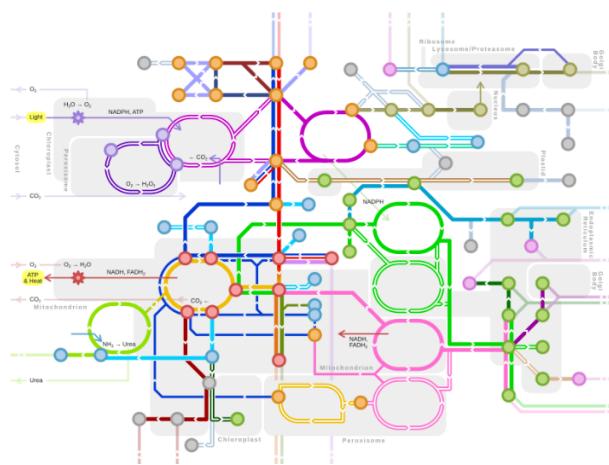
Single cell Networks



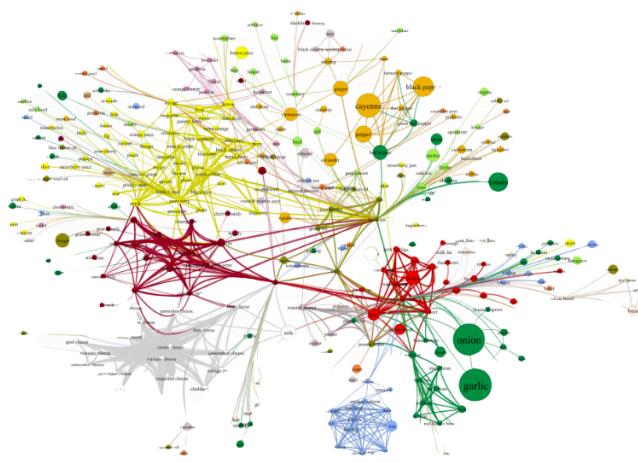
Disease Networks



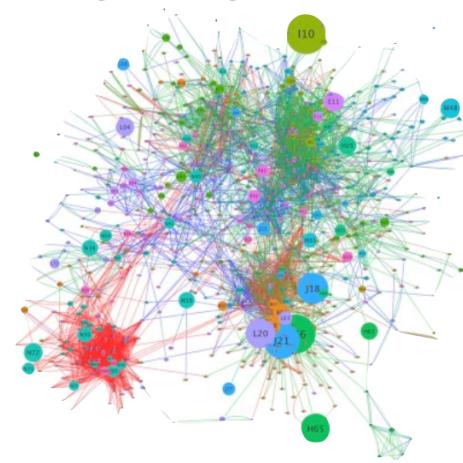
Metabolic Networks



Food Networks



Diagnosis Progression Networks



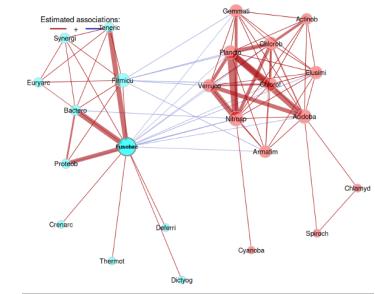
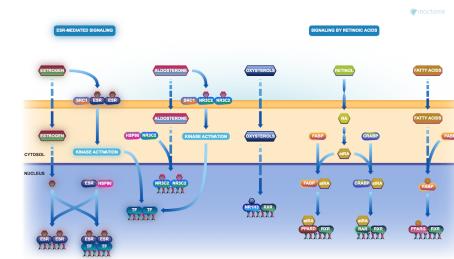
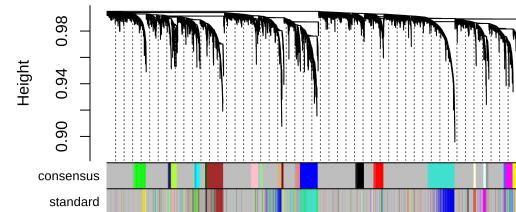
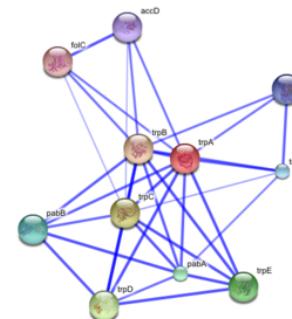
<https://towardsdatascience.com/umap-for-data-integration-50b5cfa4cdcd>
<http://snap.stanford.edu/deepnetbio-ismb/ipynb/Human+Disease+Network.html>
<https://cytoscape.org/cytoscape-tutorials/presentations/ppi-tools1-2017-mpi.html#/>
https://en.wikipedia.org/wiki/Metabolic_network
<https://www.scienceandfood.org/the-flavor-network/>

How to Build a Graph

Data to Graph

- **Data sources**

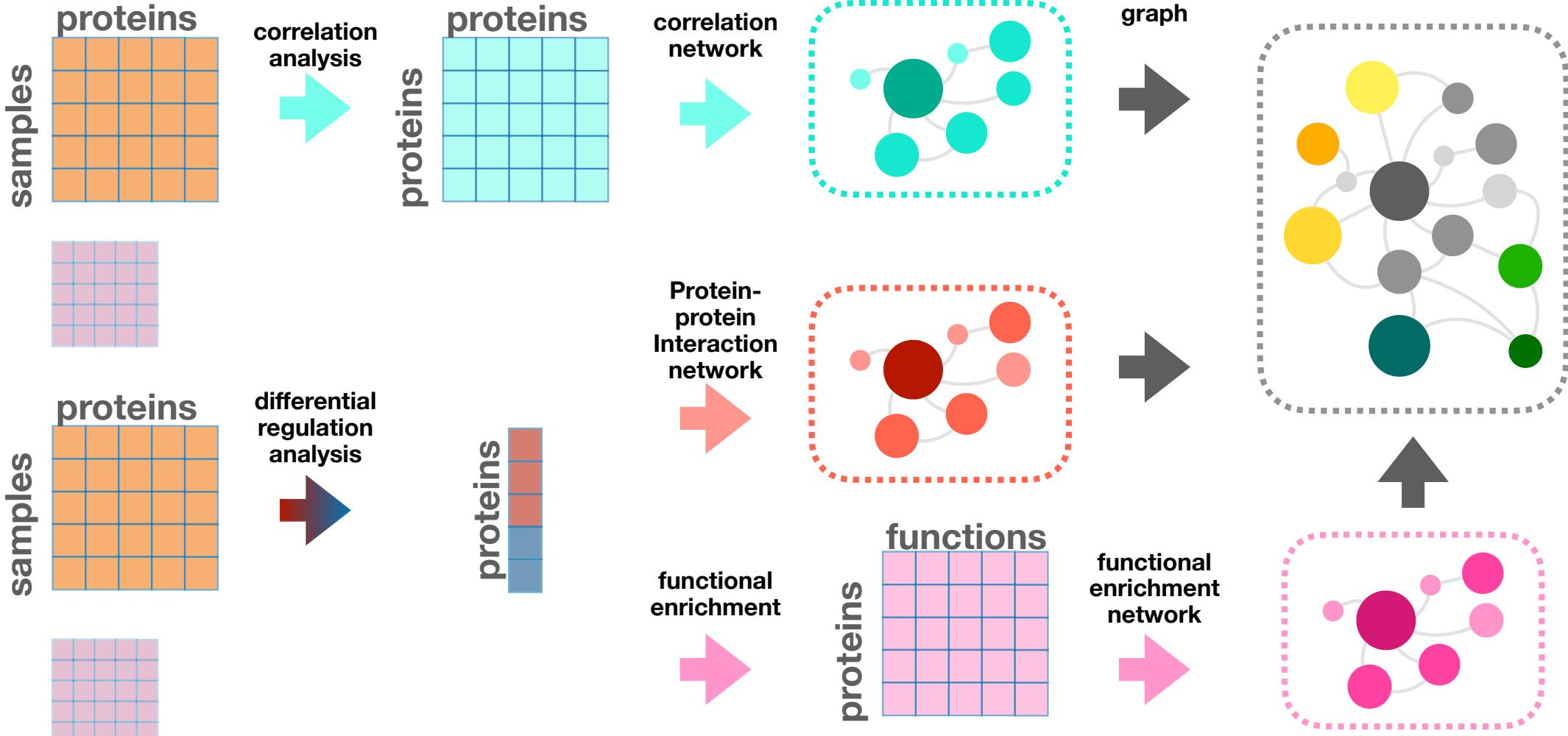
- STRING – <https://string-db.org/>
- BioGRID – <https://thebiogrid.org/>
- IntAct – <https://www.ebi.ac.uk/intact>
- REACTOME – <https://reactome.org/>
- KEGG – <https://www.genome.jp/kegg/>
- MINT – <https://mint.bio.uniroma2.it/>



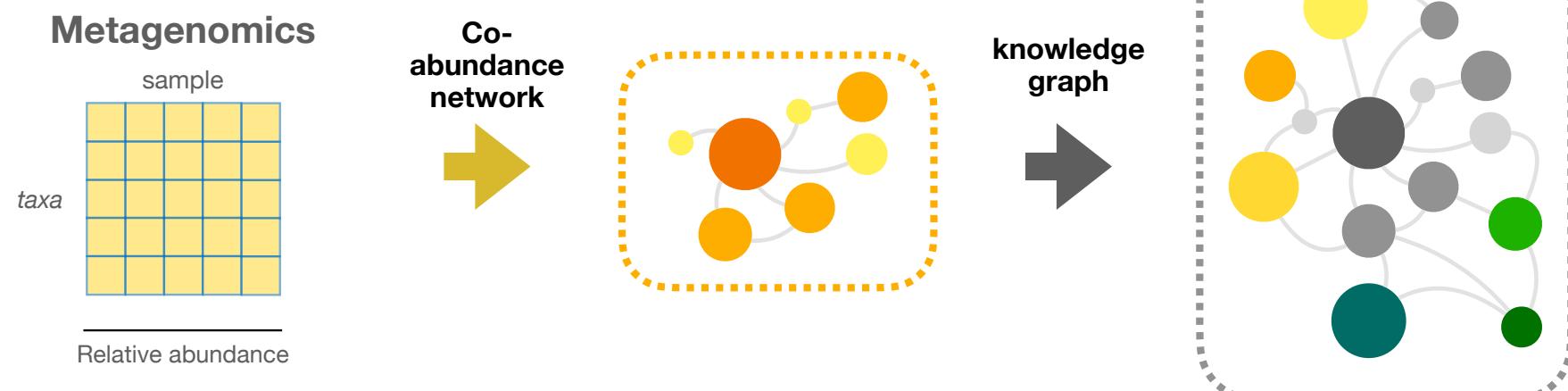
- **Correlation-based networks** – constructed by calculating pairwise correlations between entities based on their expression profiles across multiple conditions, time points, or samples (Weighted gene co-expression network analysis (WGCNA), co-abundance networks)
- **Knowledge-base approaches** – also called knowledge graphs and built by integrating heterogeneous data from multiple sources → **Knowledge Graphs**

Omics Graphs

Starting point



Microbial Association Networks – MANs

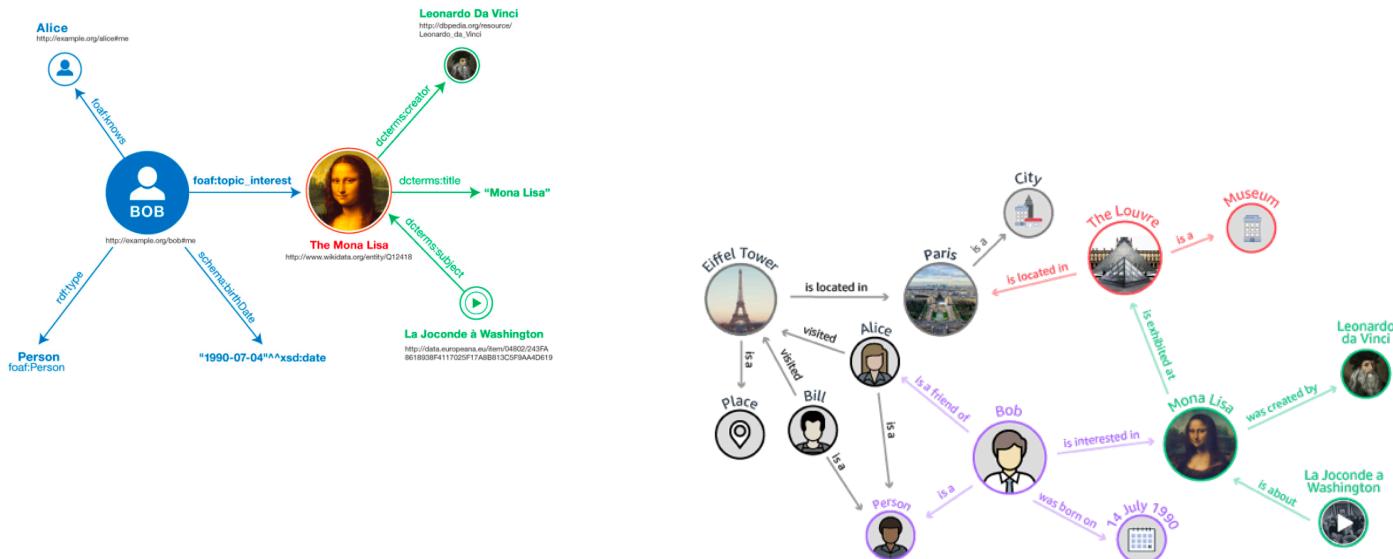


Knowledge Graphs

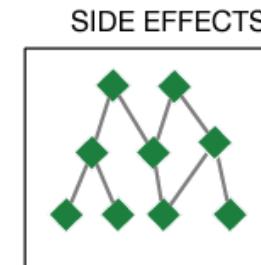
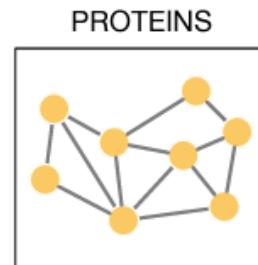
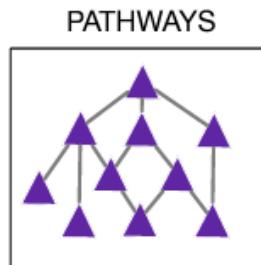
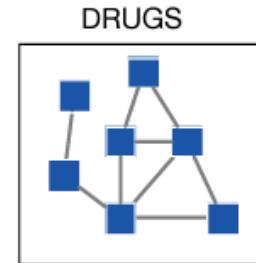
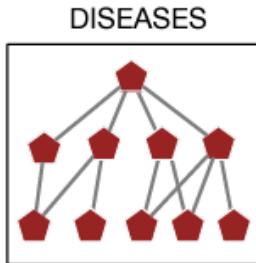
What is a Knowledge Graph (KG)

Relationships firsts everything else second

- A way to organise **knowledge/information** by defining **associations or relationships**
- These relationships facilitate the **integration, management and enrichment** of data
- The **objectives** when setting up a KG:
 - Standardisation
 - Quality
 - Reusability
 - Interpretability
 - Automation
 - Representation/Visualisation

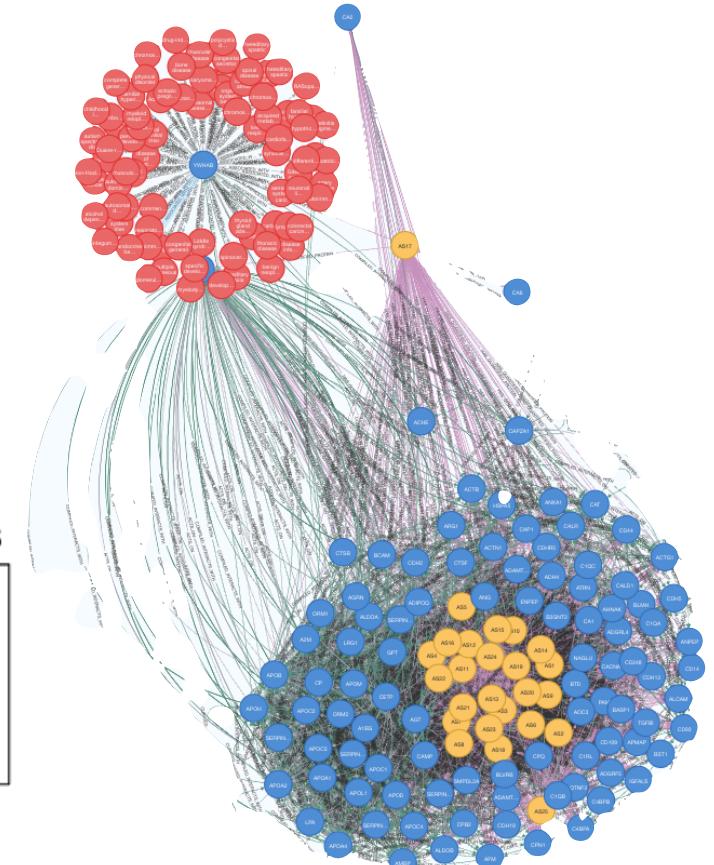
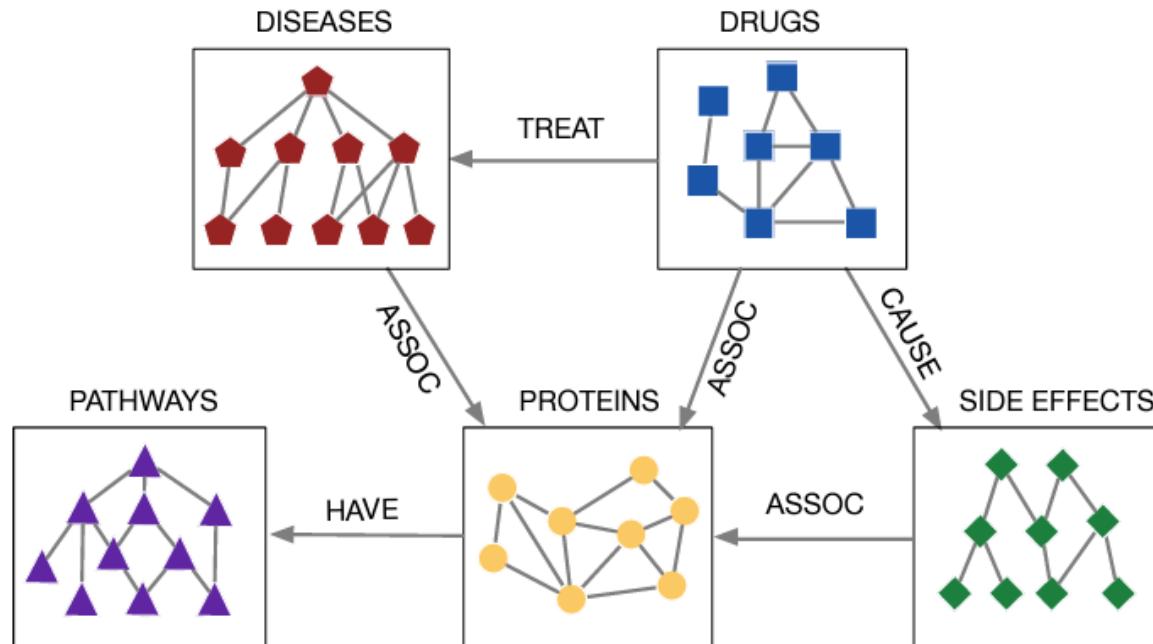


Knowledge Graph vs Graph Database



Knowledge Graph vs Graph Database

Focus on **data integration** to represent complex biological systems and be able to reason over them



Building a Knowledge Graph

1. Define the **questions** you want to answer
2. Define **what data** can be used to answer these questions and **how it is linked**
 - Data model
3. Find **where to get these data**
4. Get the data, **standardise** it and **format** it
5. Generate the **graph**
6. **Query the graph** to answer the questions

Building a Knowledge Graph

Building a Knowledge Graph

Exercise

Create a data model that allows us to answer the question:

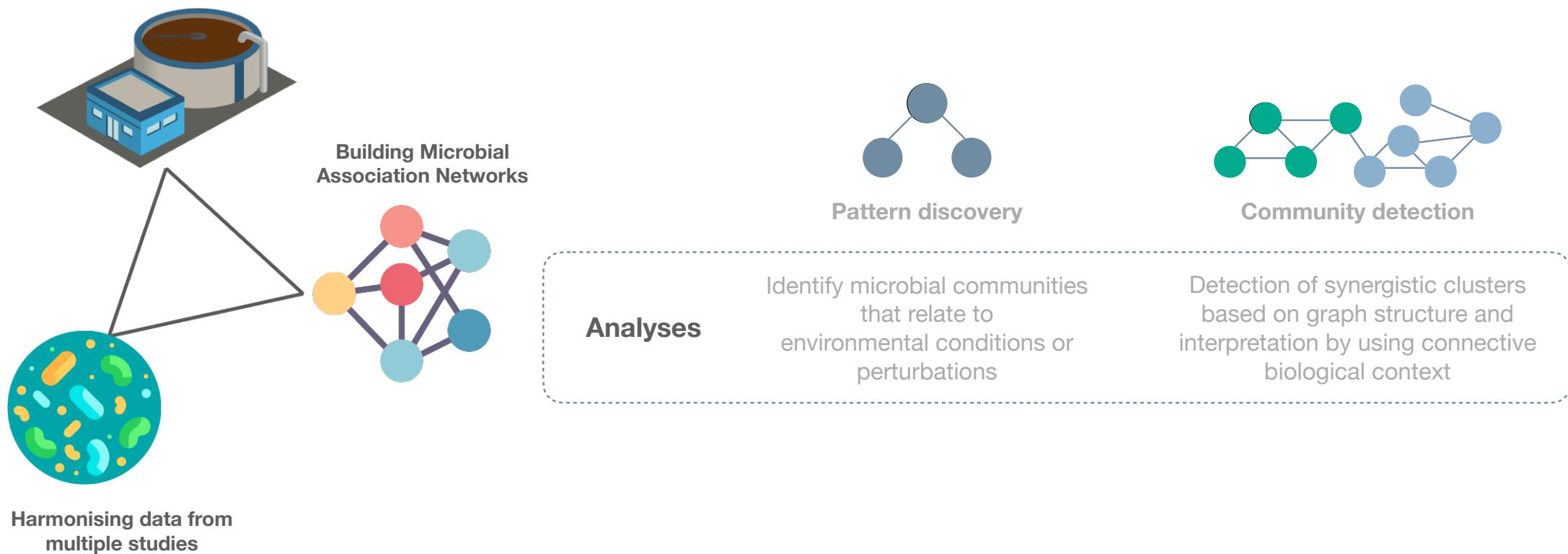
What drugs related to our disease of interest target some of the proteins identified in our experiment or relevant protein complexes and pathways?

Knowledge Graph Use Case

Wastewater Microbial Communities

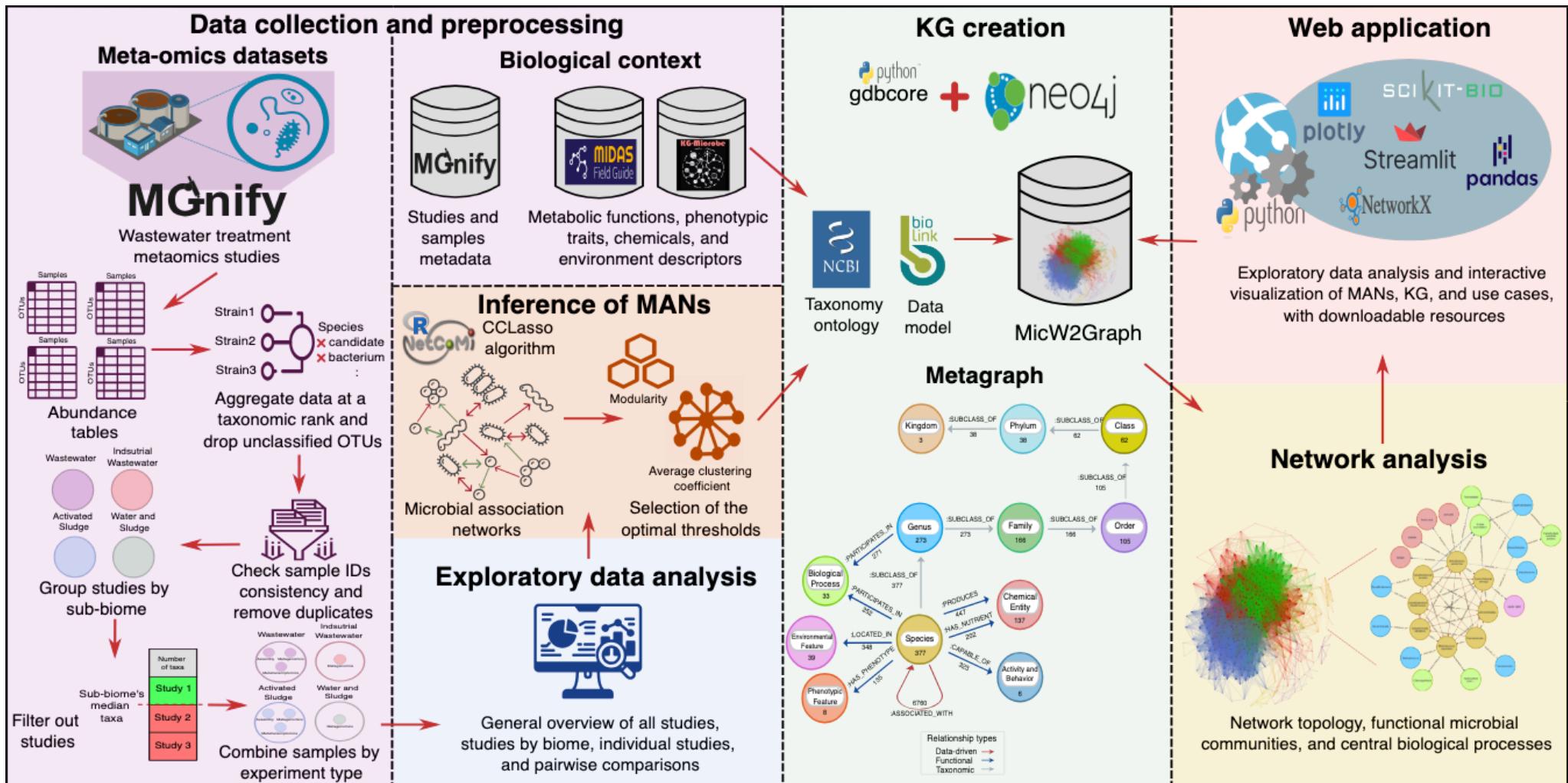


Sebastian Ayala Ruano



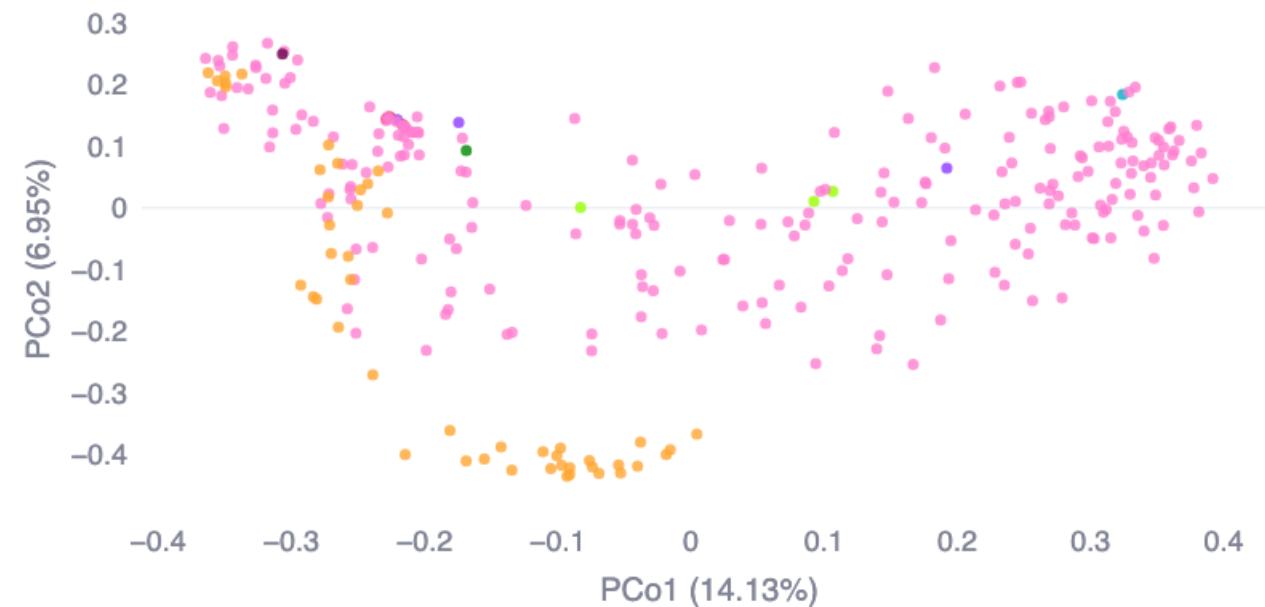
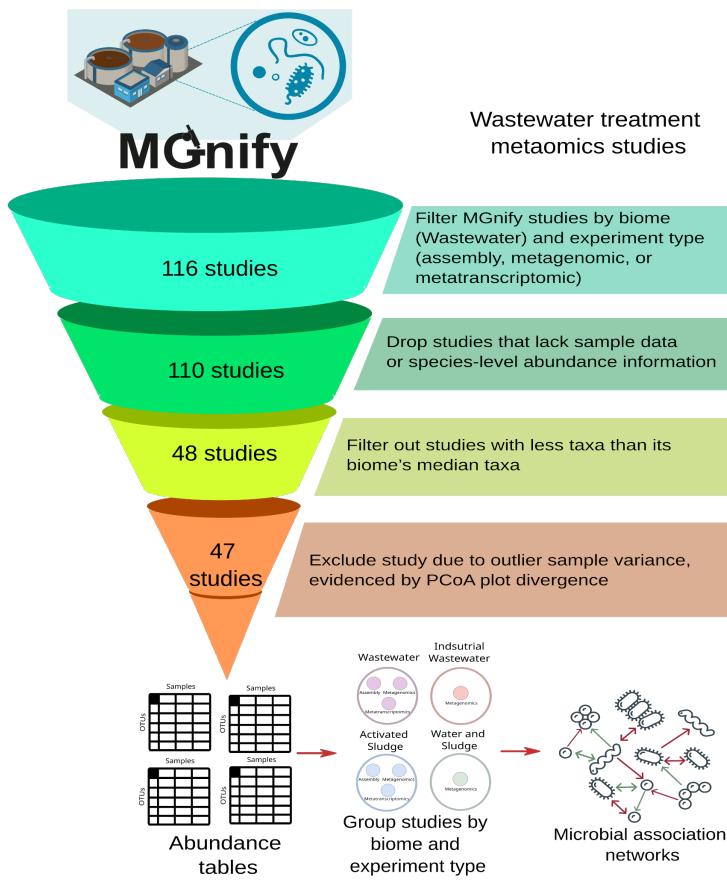
Knowledge Graph Use Case

Wastewater Microbial Communities



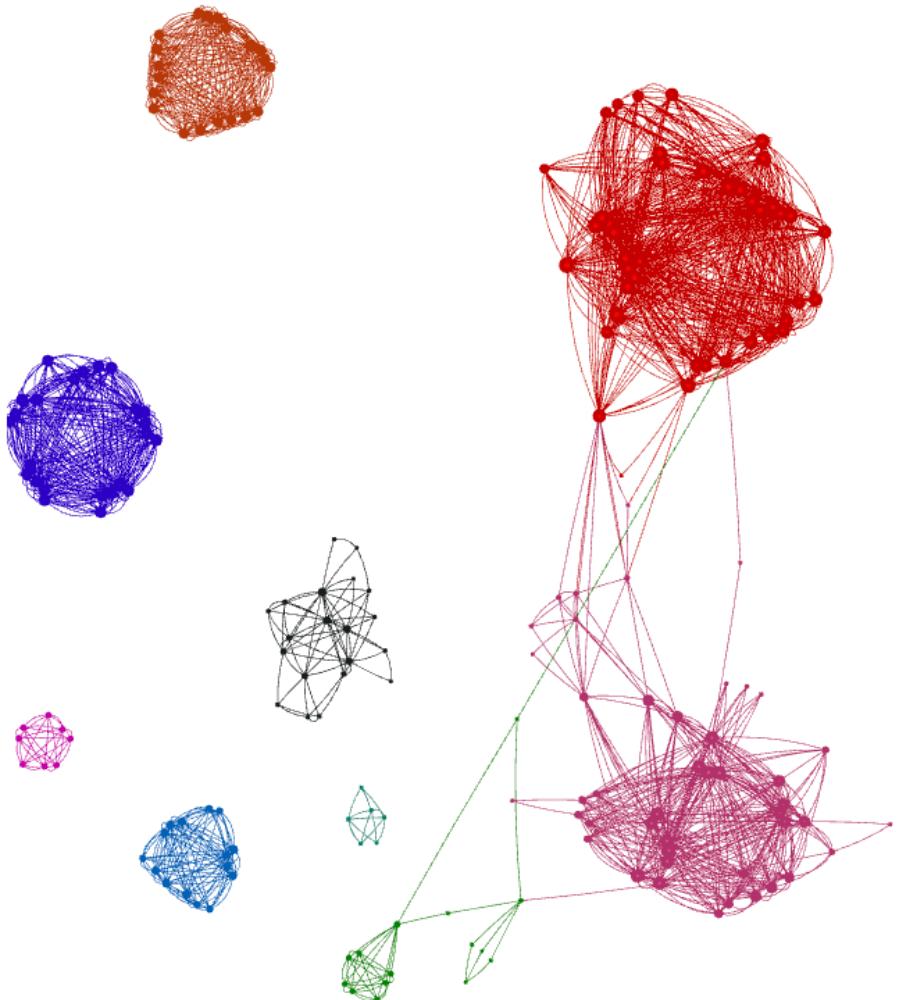
Knowledge Graph Use Case

Wastewater Microbial Communities



- thermophilic anaerobic methanogenic reactor - water
- thermophilic anaerobic methanogenic reactor - biofilm
- Wastewater treatment plant - Sewage
- Activated Sludge - Reactor
- mesophilic anaerobic reactor - wastewater
- Wastewater
- closed biogester - sludge
- wastewater treatment plant - water

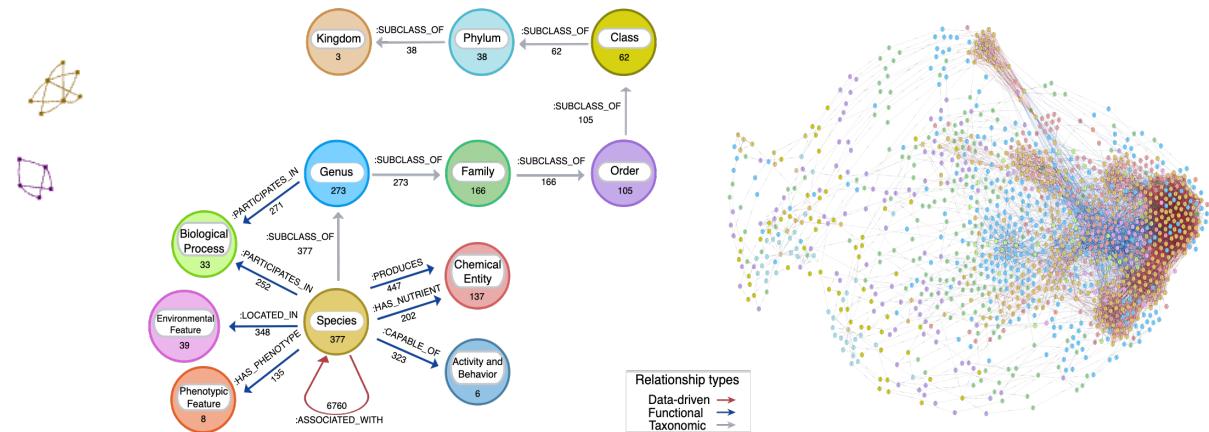
Microbial Association Networks



Microbial Association Network built using CCLASSO method — co-abundance.

Colors indicate microbial communities identified using Louvain clustering.

Functional Annotation:
Using MIDAS and KG Microbe



NetCoMi: network construction and comparison for
microbiome data in R

Stefanie Peschel , Christian L Müller, Erika von Mutius, Anne-Laure Boulesteix,
Martin Depner



Home

Exploratory data analysis

* Microbial association networks

Knowledge graph

Case studies

MicW2Graph

Building a knowledge graph of the wastewater treatment microbiome and its biological context

Microbial association networks by wastewater treatment sub-biome and experiment types

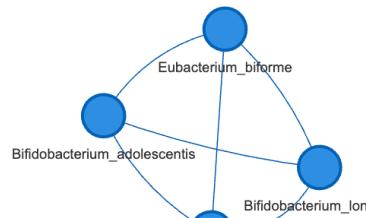
Water and sludge

Metagenomics

Number of nodes: 9

Number of relationships: 14

Add panel to control layout

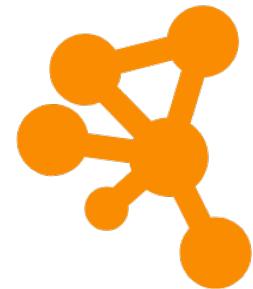


<https://micw2graph.streamlit.app/>

Tools

Cytoscape

<https://cytoscape.org/>



- An open source **software platform** for **visualising and analysing complex networks**
- Used for any kind of networks but **specialised on biological domains**:
e.g, Molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data.
- **Additional features** are available as freely available **Apps** (<https://apps.cytoscape.org/>)



Gephi

<https://gephi.org/>

- Open source **platform** for **graph visualisation** and **analysis**
- Used across **many domains** and by **many industries**
- **Performs well** for large graphs
- **Connects to Graph databases** such as Neo4j
- Some of the **visualisations and analysis** require **plugins**



Programming



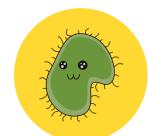
<https://networkx.org/>



<https://r.igraph.org/>

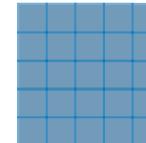
Open Source Libraries

Omics Data

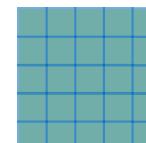


samples

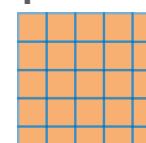
genes



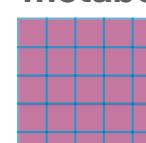
transcripts



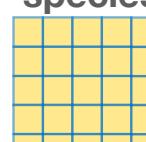
proteins



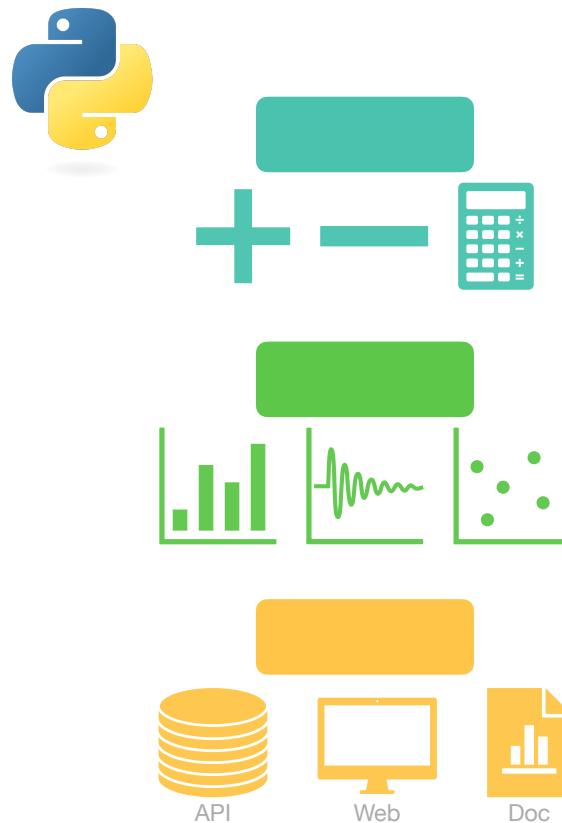
metabolites



species



Open Source Libraries



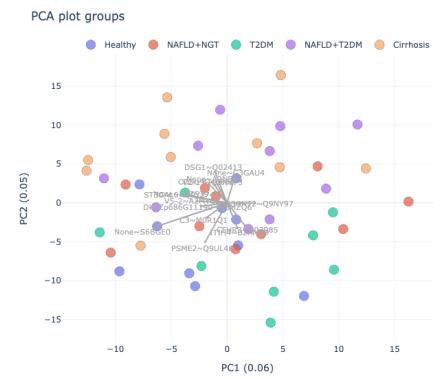
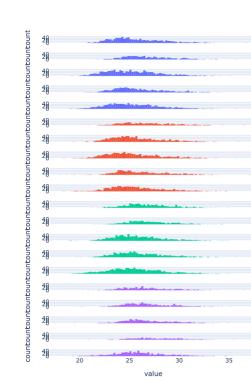
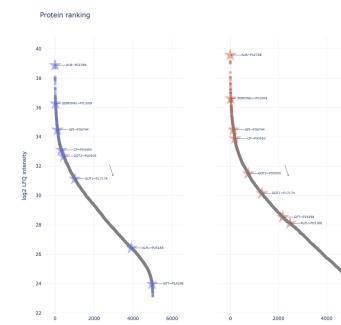
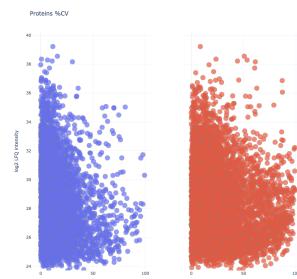
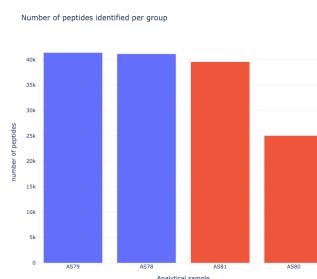
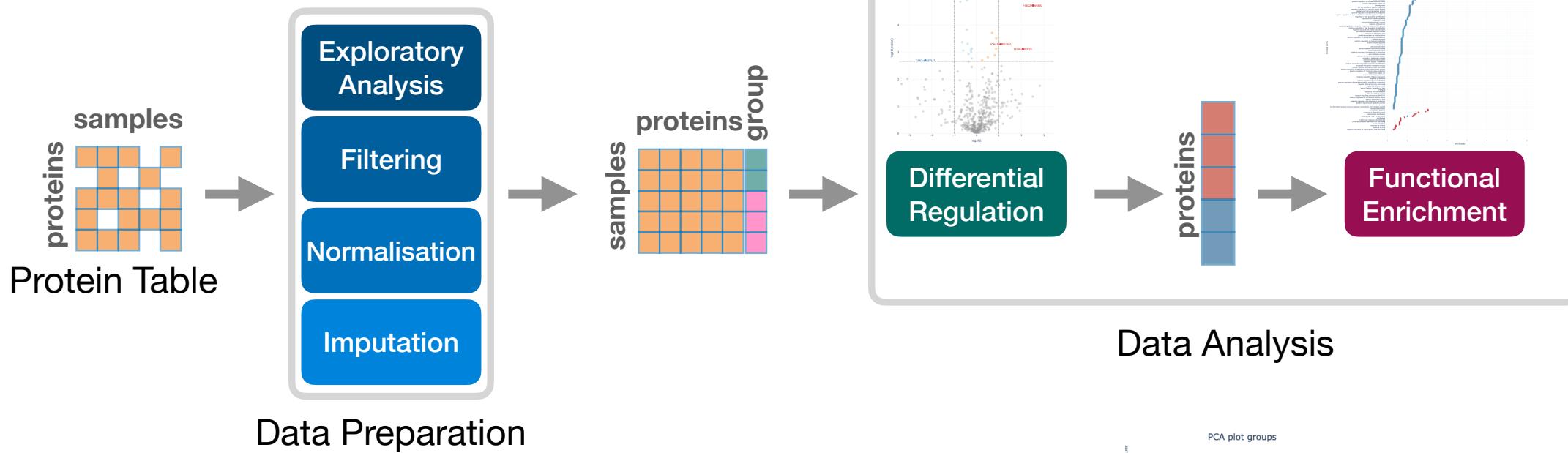
Analytics core library

<https://analytics-core.readthedocs.io/>

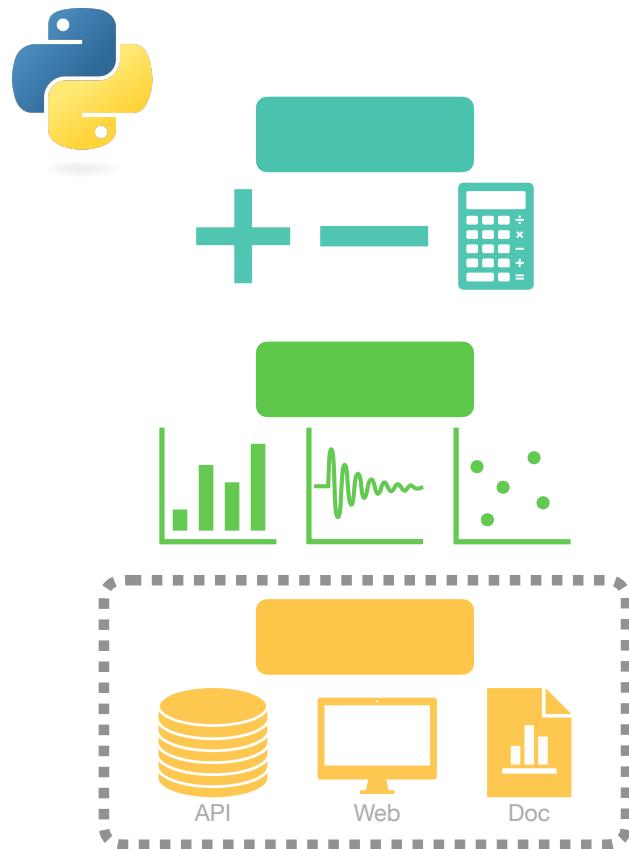
Visualization core library

<https://github.com/Multiomics-Analytics-Group/vuecore>

Analytical Workflow



Open Source Libraries



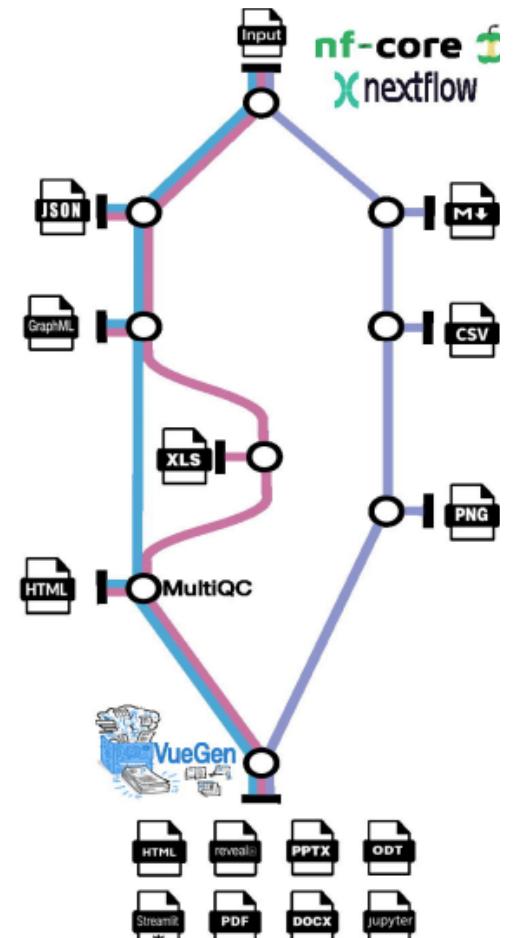
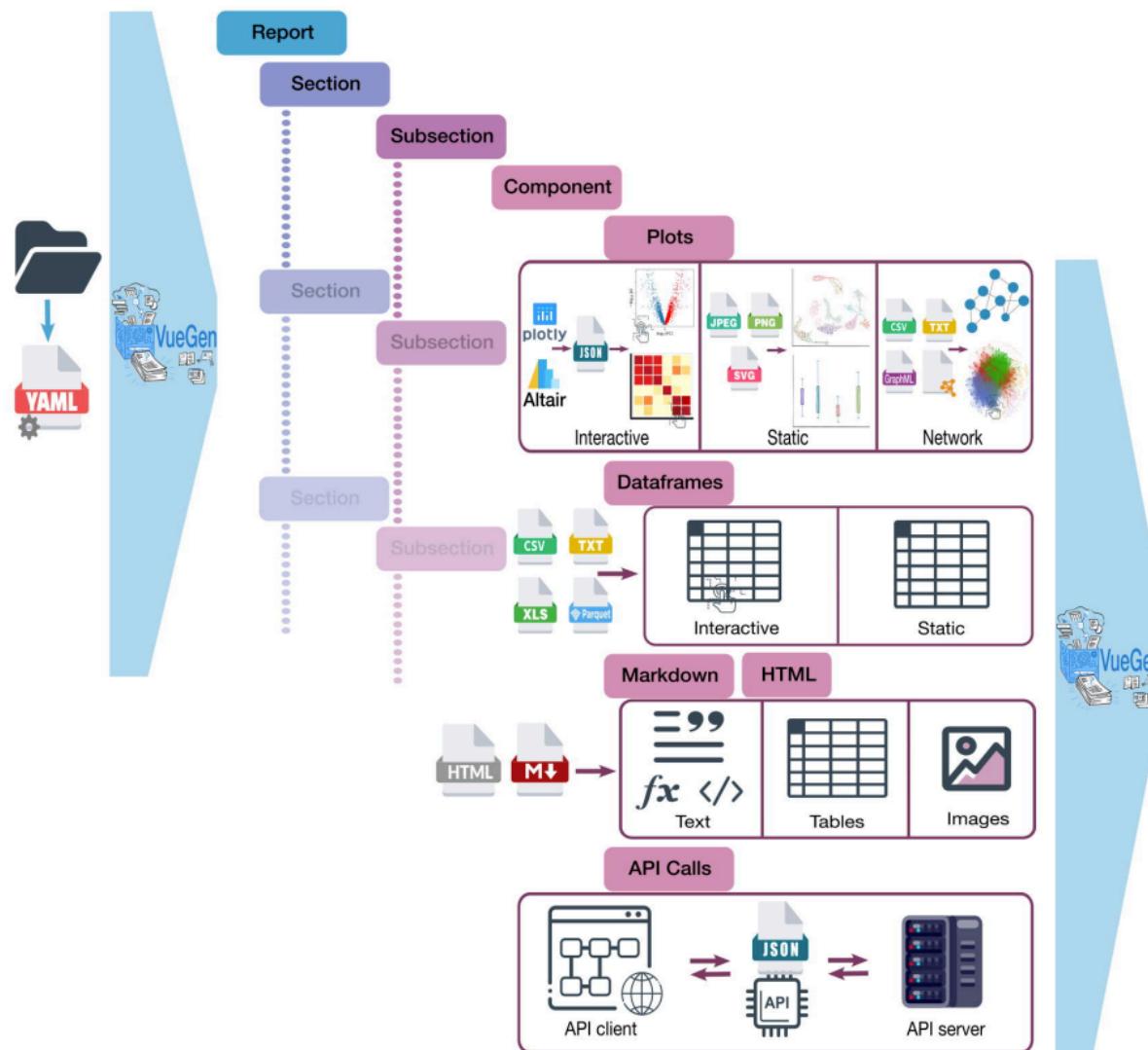
**Analytics core
library**

<https://analytics-core.readthedocs.io/>

**Visualization core
library**

<https://github.com/Multiomics-Analytics-Group/vuecore>

VueGen



VueGen: Automating the generation of scientific reports

Sebastian Ayala-Ruano, Henry Webel, Alberto Santos

doi: <https://doi.org/10.1101/2025.03.05.641152>

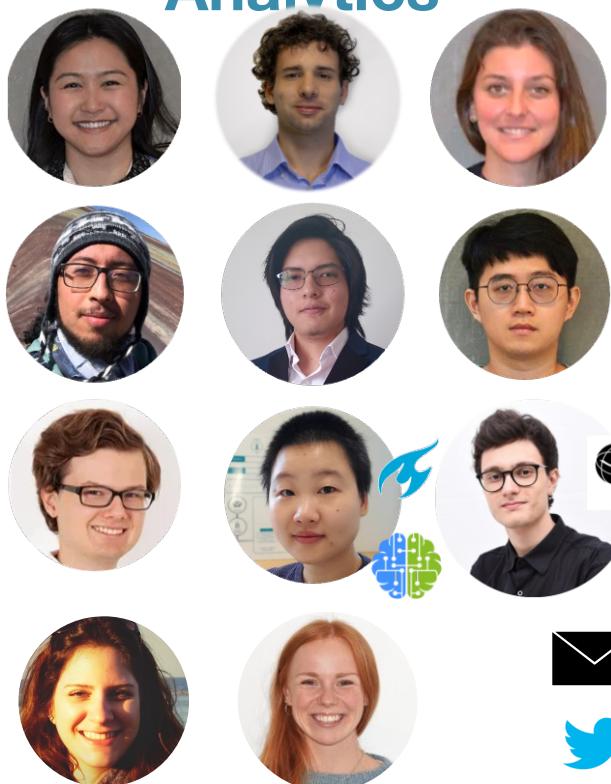
This article is a preprint and has not been certified by peer review [what does this mean?].

<https://www.biorxiv.org/content/10.1101/2025.03.05.641152v1>

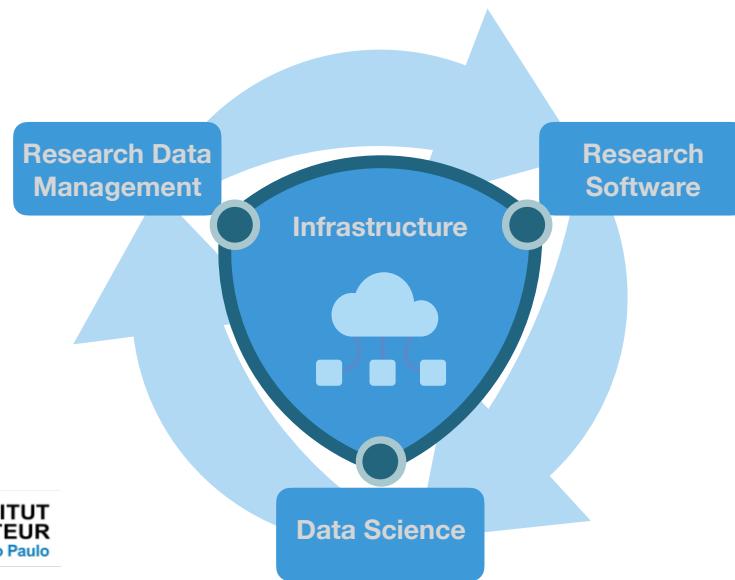
QA

Acknowledgements

Multi-omics Network Analytics



Informatics Platform



✉ albsad@dtu.dk
🐦 @albsantosdel
/github/ <https://github.com/Multiomics-Analytics-Group>
🌐 <https://multiomics-analytics-group.github.io/>

Thank you

